Using Clinical Explainable Machine Learning Models for Early Identification of Patients at Risk of Hospital-Acquired Infections

Jakobsen, Rune Sejer

DOI (link to publication from Publisher):
10.54337/aau614555530

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

USING CLINICAL EXPLAINABLE MACHINE LEARNING MODELS FOR EARLY IDENTIFICATION OF PATIENTS AT RISK OF HOSPITAL-ACQUIRED INFECTIONS

BY
RUNE SEJER JAKOBSEN

DISSERTATION SUBMITTED 2023

AALBORG UNIVERSITY
DENMARK

Using Clinical Explainable Machine Learning Models for Early Identification of Patients at Risk of Hospital-Acquired Infections

by

Rune Sejer Jakobsen

AALBORG UNIVERSITY
DENMARK

Dissertation submitted

.

Dissertation submitted:	August 2023
PhD supervisor::	Clinical Professor Peter Leutscher, Aalborg University
Co- Ph.D. supervisor:	Associate Professor Thomas Dyhre Nielsen, Aalborg University
	MD, Ph.D., Kristoffer Koch, Aalborg University Hospital
PhD committee:	Professor Stephen Rees (chair) Aalborg University, Denmark
	Professor Joseph Randall Moorman University of Virginia, USA
	Associate Professor Simon Tilma Vistisen Aarhus University, Denmark
PhD Series:	Faculty of Medicine, Aalborg University
Department:	Department of Clinical Medicine

ISSN (online): 2246-1302
ISBN (online): 978-87-7573-645-4

Published by:
Aalborg University Press
Kroghstræde 3
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Stibo Complete, 2023

PROFILE

I grew up in a small town called Hørning, just between Aarhus and Skanderborg, mostly known for its annual festival in the woods. Being only four years of age, I was diagnosed with diabetes mellitus type 1, which perhaps unawarely sparked my earliest interest in healthcare, that has later become an integral part of my life. I was blessed with great teachers in the local primary school that inspired my interest in math, physics, and human biology, which came naturally to me. I later attended high school in Skanderborg to pursue a growing passion for natural sciences. However, high school also taught me a hard-won and valuable lesson from a scattering discipline: *"hard work beats talent when talent does not work hard"*.

My discipline was later refined after a gap year, including responsible full-time employment at the local supermarket and a two-month stay in Xiamen, China, where I experienced an entirely different work ethic than Danish high school standards. Foremost, the gap year allowed me to gain perspective on my passions, making my curiosity return to inspirational sources of healthcare, math, and human biology. I, therefore, applied for Biomedical Engineering and Informatics at Aalborg University.

I got my Bachelor's degree in Biomedical Engineering and Informatics from Aalborg University, which also shaped my earliest academic interest in challenging the status quo at a higher level. In my Master's in Biomedical Engineering and Informatics, I gained a deeper understanding of how applied math could result in better decisions from algorithmic thinking – and why healthcare may benefit from solutions relying on, e.g., machine learning.
In the last four years of my life, I have been employed at the Centre for Clinical Research at the Regional Hospital in Hjørring. I have been involved in several projects concerning machine learning in healthcare, foremost the pioneering work related to my Ph.D. about machine learning and hospital-acquired infections, which included a conspicuous collaboration between the regional Business Intelligence Unit, the Regional Hospital in Hjørring, Aalborg University hospital, and Aalborg University. In my experiences from the academic disciplines, I learned that commitment to excellence has an additional dimension to the hard-won and valuable lesson learned after high school: *"working right beats hard work when hard work is not done right"*. I firmly believe that this lesson invites a talk about the role of machine learning in the future of healthcare.

PREFACE

In the eyes of the modern world, the flaming hype of *machine learning* and *artificial intelligence* appears comprehensive. Naturally, these firey sparks have also reached the doorstep of healthcare and infection control. Can we, e.g., predict and avoid an adverse event before it would have occurred? Can we assist in the surveillance of risk factors, provide decision support for optimal clinical therapy, and increase the chances of survival and well-being? Can we, by all means, aim for a better outcome in tomorrow's infection control than we were able to reach yesterday? The promises of machine learning may invite a solution to these questions. However, they somehow struggle to find their way into the clinical routine. As a result, their potentials remain unresolved – but also, in many cases, remarkably unexplored.

I was introduced to machine learning in my master's program at Aalborg University under a different wording, namely *pattern recognition*, which may reveal my background as a Biomedical engineer. However, my academic journey with hospital-acquired infections began later at the Centre for Clinical Research, where I worked alongside other researchers and clinical experts, who also undoubtedly shaped the perspectives gained in the research field. For instance, I learned that the bridge between machine learning and healthcare is, first and foremost, reliant on the perspectives of establishing trust. How can the patient trust that adverse outcomes are avoided? How can the clinician trust in being equipped with the best tools and trust the assistance provided by such? How can the organizational decision-maker trust that therapy does not exceed the expected cost? These perspectives inspired my work to explore how machine learning models could be developed to control and manage hospital-acquired infections and how they may fit the daily clinical routine.

Hopefully, after you have read this thesis, you will feel inspired by the perspectives and ideas we propose. Perhaps you will also better understand the hype of machine learning for the prevention of hospital-acquired infections. While the past four years have been a privilege, they have not been without challenges, choices, and compromises, but every bit of it has made it a journey for which I am truly grateful. Overall, I have, in my own opinion, matured into an independent researcher in the cross-disciplinary field of biomedical engineering. In this process, I also gained insight, and the most profound respect, into the crossed field of medicine, particularly hospital-acquired infection control, and data science, particularly machine learning.

The Ph.D. dissertation kindly received funding from the North Denmark Regional Hospital, the Regional Health Innovation Pool (Regionens sundhedsinnovationspulje), Marie Jensen og Jensine Heibergs Foundation, and Niels Jensens Foundation. This study would not have been possible without this assistance.

ACKNOWLEDGEMENTS

A Ph.D. dissertation is hardly a single person's work. Without the support of family, colleagues, and friends, completing the work of the thesis might not have been possible.

First and foremost, I would like to acknowledge my girlfriend and fiancee, Emma Rosendal Jensen, for her outstanding patience, love, and tolerance. Knowing me, it must not always have been easy for you – especially when I was more dedicated to my work than investing in our relationship. However, when I devoted my time elsewhere, you held the beacon of loving light for both of us with great care. That you ever said *yes* to my proposal in France during the Ph.D.-period is beyond a level of luck I could ever imagine.

The thesis has been carried out under excellent and ambitious supervision by *the three wise men*: Peter Leutscher, Kristoffer Koch, and Thomas Dyhre Nielsen. My deepest gratitude goes to you for our academic discussions and the time you reserved for engaging in all the many debates with very different natures. While I surely was not always the easiest Ph.D. fellow to supervise, you never lost faith in me or our work – also when the ideas became too many and my stubbornness too massive. You also engaged in my personal life, which was beyond what I expected but only could have hoped for. To me, you demonstrated inspiring ambitions on my behalf and never held back on a follow-up or offering a guiding talk. While I have learned so much from you, I hope my gratitude finds you well and builds upon the confidence that you most certainly deserve. It has truly been an honor! Hopefully, there will be room for collaboration in the future.

I would also like to thank my current and previous colleagues, who made the journey toward the completion of this thesis spectacular. Working at the Centre for Clinical Research, North Denmark Regional Hospital, I was invited to all kinds of funny, clinical, and even philosophical discussions. For instance, within the first week of employment as only a research assistant, I was challenged with the great question: *"Artificial intelligence, you say… can you please tell us; what is normal intelligence?"…* Like a deer caught in the headlights…

I offer to Andreas Egmose, Thomas Mulvad, Mads Nibe, Jimmy Klitgaard, and Maj Skaarhøj, whom I, among many more, consider my colleagues at the Regional Business Intelligence unit, a warm thank you. We collaborated on all kinds of challenges in the great arena of data utilization, from which I believe we have all learned a lot – also about the much more work that needs to be done in marrying machine learning and healthcare. While also building collaborations and a machine learning team as a part of the BI office, you also provided excellent consultancies on the performed data management from the big, yet messy, amount of data available in your warehouse. Thank you!

I was also lucky to find inspiration and kindly learn from the experiences and talks with professor emeritus Steen Andreasen, who also functioned as the opponent in my pre-defense with the result of very constructive feedback. Likewise, I am also grateful to professor Henrik Nielsen and the Department of Infectious Diseases at Aalborg University Hospital. Besides my visit, where I observed your daily clinical routine, I had several honors of discussing our ideas of machine learning in control of hospital-acquired infections with your team afterward, e.g., at the NCSMID conference.

Thorough and humble gratitude also goes to family and friends. Being almost forced to hear my tales and stories must have been exhausting, but you never bailed out of my tellings. I offer special thanks to Emil Riis Hansen, my lifelong friend and co-Ph.D. fellow, for all the professional, philosophical, and deeply personal talks along the way. I believe your insight helped me understand our ideas and transform the co-identified challenges of the world into dreams for the future. Let us continue to have these inspiring talks! Special gratitude also goes to my father, Henning Sejer Jakobsen, with whom I share academic ambition. An appetite for knowledge, that resulted in collaboration with you and Aarhus University, which further led to submitting our work with artificial intelligence in strategic idea portfolios and radical innovation. My stubbornness may not be an all-self-grown quality, but neither is my academic curiosity and willingness to challenge the status quo. Thank you!

While I cannot mention all, I still hope that no one feels left out just because you are not mentioned in this acknowledgment. You are not forgotten! A humble and most sincere *thank you* to all who have been a part of the journey toward completing the ph.d. thesis! It has been a privilege.

ENGLISH SUMMARY

Background: Hospital-acquired infection (HAI) is a common, yet often preventable, complication during hospital admission. They are associated with higher mortality rates, prolonged lengths of stay, increased hospital expenditures, and considerable patient discomfort. Traditional strategies for HAI rely on, e.g., timely preventive measures, thorough surveillance, careful usage of indwelling catheters, and rational use of antibiotic therapy. However, recent studies report increased incidences, and with the emergence of multi-resistant pathogens, negative consequences of HAI will likely remain, or even increase, in the coming years.
Machine learning (ML) models for early identification of patients at risk of HAI are suggested to enable timely and targeted preventive strategies for patients in most need. With these perspectives, chances of avoiding HAI in the future may increase with a resulting advancement to decreased incidence levels – if we can predict them in the first place. However, clinicians are often challenged in the interpretation of the predictive outcomes provided by the ML models, which may negatively impact the adoption into the clinical routine.

Aim: To explore how ML models can be developed for risk stratification within 24h of admission for HAI and how these may fit the clinical routine. Two studies target hospital-acquired urinary tract infections (HA-UTIs), and one targets hospital-acquired bacteremia (HAB). Study I further elucidated model-specific clinical explainability perspectives, whereto Study II explored model-agnostic clinical explainable methods. Study III addressed feasibility, including a health-economic cost-effective analysis, in ML-guided routine replacement of peripheral venous catheters (PVC) to avoid adverse HAB instances.

Materials and methods: The studies investigated patient data representing 138.560 hospital admissions in the North Denmark Region from 01.01.2017 to 31.12.2018. 51 health socio-demographic- and clinical features were extracted for the purpose. The case definitions of HA-UTI and HAB relied on a national surveillance system to monitor the frequency of occurrences, called the Healthcare-Associated Infection Database (HAIBA).
For the first study, Bayesian networks are developed on a full- (51 features) and reduced dataset (5 features), selected from a combination of expert knowledge and machine suggestions. Four methods are applied to compare different structure learning approaches: manual expert knowledge, naïve, automatic tree-augmented-naïve- and Peter-Clark-algorithm. The expectation-maximization algorithm is applied for parameter learning.
In the second study, seven different ML models are developed over a full- (51 features) and two reduced datasets (10 features), selected from an automated χ^2 test and manual expert knowledge, respectively. The seven different ML models are logistic regression, neural network, Bayesian network, decision tree, random forest, gradient boosting, and AdaBoost. The Shapley Additive explanation (SHAP) method

is applied to support model-agnostic explainability.
In the third study for predicting the risk of HAB, five different ML models are applied to a dataset with 48 categorical and continuous features. The five ML models are logistic regression, neural network, Bayesian network, decision tree, and random forest. The numbers needed to harm (NNH) and an incremental cost-effectiveness ratio (ICER) is used to address the feasibility of the ML-guided routine replacement strategy of PVC for the patients in most need.

Findings: The first study revealed that the structures developed from expert knowledge are associated with better domain representation and reached the best area under the curve (AUC) of 0.746. Bayesian networks handle missing data well, rely on probabilistic reasoning, and enable inspection of the evidence and dependencies, which can be associated with supporting model-specific clinical explainability for early risk prediction of HA-UTI.
The second study brought to light that an automated χ^2 test favors laboratory results in its feature selection, where manual expert knowledge has a better knowledge representation with a mix of health-sociodemographics and clinical features. The best-performing ML model is a neural network based on the full dataset, reaching an AUC of 0.758. Model-agnostic explainability is demonstrated with a SHAP summary- and forceplot to how the model weights the features over the population and explain why a model reaches a given result for a single risk prediction of HA-UTI.
Lastly, the third study revealed that a random forest model reached an AUC of 0.82 in risk-stratifying HAB. From a decision threshold associated with providing routine replacement of PVC for 20 % of the cohort, an associated NNH is 766 and an ICER of DKK 1,440,495.00 per avoidable HAB-related death is reported. Lowering the decision threshold in the ML models will also lower the NNH, but with a resulting increase in the ICER and vice versa from an increase of the decision threshold.

Conclusions: The results indicate that explainable ML can predict HAI risk within 24h of admission, enabling targeted preventive measures and efficient future HAI control strategies for patients in most need. The dissertation shows that varied health and clinical data can be effectively used in ML models for these predictions. The feasibility of ML-guided strategies, e.g., PVC replacement, can be evaluated through cost-effectiveness analysis. The developed ML models support model-specific and model-agnostic clinical explainability, and may be considered feasible, potentially fitting in the era of responsible ML in healthcare.

DANSK RESUME

Baggrund: Hospitalserhvervede infektioner (HI) er almindelige, men ofte forebyggelig, komplikationer under hospitalsindlæggelse. De er forbundet med forhøjet dødelighed, forlængede hospitalsophold, øget hospitalsudgifter og er tilmed til stor ubehag for patienterne. Traditionelle strategier mod HI inkluderer f.eks. rettidig forebyggende foranstaltninger, grundig overvågning, forsigtig brug af indlagte katetre og rationel brug af antibiotika. Nylige danske studier har dog rapporteret en stadig stigende forekomst af HI, hvortil udfordringer med multiresistente patogener også bidrager til et øget pres infektionskontrollen. Som følge heraf er det sandsynligt at forvente, at de negative konsekvenser fra et forløb med HI vil forblive, eller måske endda stige, i de kommende år.
Modeller kendt fra maskinlæring (ML) til tidlig genkendelse af patienter med risiko for HI kan måske muliggøre rettidig omhu og målrettet forebyggelse for patienter med størst behov – hvis vi kan forudsige og finde dem i første omgang. Klinikere er dog ofte udfordret i fortolkningen af modellerne og deres prædiktive resultater, hvilket medfører, at chancerne for adoptionen i den kliniske hverdag falder drastisk. Denne iagtagelse er ikke ny og har ledt til voksende evidens og retningslinjer for fremtidige perspektiver for ML sundhedsvæsenet, der efterspørger, at udviklingen af kliniske ML-modeller i højere grad tilstræber at inkludere klinisk ekspertviden i den overordnede udviklingsproces, samt tilbyder en forklaring på deres ræsonnement i de kliniske forudsigelser.

Formål: At undersøge, hvordan forskellige ML-modeller kan udvikles til risikostratificering inden for 24 timer efter indlæggelse for HI, og hvordan disse kan passe til den kliniske hverdag. To studier er rettet mod hospitalserhvervede urinvejsinfektioner (HUVI) og et studie er målrettet hospitalserhvervet bakteriæmi (HB). Studie I belyser yderligere modelspecifikke kliniske forklarbarhedsperspektiver, hvortil Studie II kaster lys over modelagnostiske kliniske forklarbare metoder. Studie III omhandler gennemførlighed, herunder en sundhedsøkonomisk omkostningseffektiv analyse, i ML-guidet rutinemæssig udskiftning af perifere venekatetre (PVK) for at undgå uønskede HB-tilfælde.

Materialer og metoder: Studierne undersøger patientdata, der repræsenterer 138.560 hospitalsindlæggelser i Region Nordjylland fra 01.01.2017 til 31.12.2018. I alt er 51 sociodemografiske og kliniske variabler inkluderet til forudsigelse af HI. Definitionerne af HUVI og HB er baseret på et nationalt overvågningssystem til monitorering af frekvenser for hændelser, kaldt Healthcare-Associated Infection Database (HAIBA).
I første studie bliver forskellige Bayesianske netværk udviklet og sammenlignet, der trænes på henholdsvis et fuldt- (51 variabler) og reduceret datasæt (5 variabler), udvalgt fra en kombination af ekspertviden og maskinforslag. Fire forskellige metoder bliver anvendt til strukturlæring over de to datasæt: manuel ekspertviden, naiv, Tree-Augmented-naive- og Peter-Clark-algoritmen. Expectation-maximization algoritmen bliver anvendt til parameterindlæring.

I andet studie bliver syv forskellige ML-modeller udviklet over et fuldt- (51 variabler) og to reducerede datasæt (10 variabler), udvalgt fra henholdsvis en automatiseret χ^2-test og manuel ekspertviden. De syv forskellige ML-modeller er logistisk regression, neuralt netværk, Bayesiansk netværk, beslutningstræ, random forest, gradient boosting og AdaBoost. Shapley Additive Explanation (SHAP) metoden bliver anvendt til at understøtte modelagnostisk forklaring.
I tredje studie bliver risikoen for HB bestemt ud fra fem forskellige ML-modeller anvendt på et datasæt med 48 kategoriske og kontinuerte variabler. De fem ML-modeller er logistisk regression, neuralt netværk, Bayesiansk netværk, beslutningstræ og random forest. Numbers-needed-to-harm (NNH) og en inkrementelt kost-effektivitet ratio (IKER) bliver brugt til at adressere gennemførligheden af ML-guidet rutinemæssige udskiftningsstrategi af perifær vene kateter (PVK) for de, procentvise del af patienterne, der har størst behov.

Resultater: Det første studie afslører, at strukturerne udviklet fra ekspertviden er forbundet med bedre domænerepræsentation og nåede den bedste area under the curve (AUC) på 0,746. Bayesianske netværk håndterer manglende data godt, tillader probabilistisk ræsonnement og muliggjorde inspektion af evidens og afhængigheder, som kan associeres med at understøtte modelspecifik klinisk forklarbarhed til tidlig risikoforudsigelse af HUVI.
Det andet studie viser, at en automatiseret χ^2-test favoriserer laboratorieresultater i dens variableudvælgelse, hvor manuel ekspertviden har en bedre videnrepræsentation med en blanding af sundheds-sociodemografi og kliniske variabler. Den bedst-performende ML-model er et neuralt netværk baseret på det fulde datasæt, der nåede en AUC på 0,758. Model-agnostisk forklarbarhed bliver demonstreret med et SHAP summary- og forceplot til, hvordan modellen vægter variablerne, herved hvorfor en model når et givet resultat for en enkelt risikoforudsigelse af HUVI.
Endeligt afslører det tredje studie, at en random forest model nåede en AUC på 0,82 i risiko stratificering af HB. Fra en beslutningstærskel forbundet med at levere rutinemæssig udskiftning af PVK for 20 % af kohorten med størst risiko for HB, medfølger en NNH på 766 og en ICER på DKK 1,440,495.00 pr. undgået HB-relateret dødsfald. Sænkning af beslutningstærsklen i ML-modellerne vil også sænke NNH, men med en deraf følgende stigning i IKER, og vice versa fra en stigning af beslutningstærsklen.

Konklusioner: Resultaterne af de tre studier indikerer, at forklarbar ML kan benyttes til forudsigelse HI-risiko inden for 24 timer efter indlæggelsen, hvilket muliggør målrettede forebyggende foranstaltninger og effektive fremtidige HI-strategier. Resultaterne viser dertil, at en blanding af sundhedsrelaterede- og kliniske data effektivt kan bruges i ML-modeller til disse forudsigelser. De udviklede ML-modeller kunne understøtte model-specifik- og model-agnostisk forklarbarhed, hvilket potentielt øger deres tilpasning til en klinisk kontekst. Dertil kunne ML-guidede strategier til PVK skift potentielt udfordre eksisterende alternativers gennemførlighed. Denne afhandlings resultater og fremgangsmåde passer muligvis ind i en æra med ansvarlig ML i sundhedsvæsenet.

TABLE OF CONTENTS

LIST OF SCIENTIFIC PAPERS

Manuscript I: Jakobsen RS, Nielsen TD, Leutscher P, Koch K, A study on the risk stratification for patients within 24 hours of admission for risk of hospital-acquired urinary tract infection using explainable Bayesian Network models, Journal of Health Informatics, 2023. (Resubmitted – minor revisions)

Manuscript II: Jakobsen RS, Nielsen TD, Leutscher P, Koch K, Clinical explainable machine learning models for early identification of patients at risk of hospital-acquired urinary tract infection, Journal of Hospital Infection, 2023; https://doi.org/10.1016/j.jhin.2023.03.017. (Published)

Manuscript III: Jakobsen RS, Nielsen TD, Christensen JJ, Sørensen SS, Leutscher P, Koch K, A machine learning guided strategy for the prevention of hospital-acquired bacteremia: a comparative cost-effectiveness analysis, Nature Partner Journals (npj) Digital Medicine, 2023. (Submitted)

TABLES AND FIGURES

Figures:

Tables:

ABBREVIATIONS

AdaBoost (AD)

American Society of Anesthesiologists (ASA) score

Antimicrobial resistance (AMR)

Area under the curve (AUC)

Artificial intelligence (AI)

Bayesian Network (BN)

Categorical and Regression Tree (CART)

Charlson Comorbidity Index (CCI)

Civil personal registration (CPR)

Clinical decision support systems (CDSS)

Community-acquired infection (CAI)

Conditional probability table (CPT)

Cross-validation (CV)

Danish crowns (DKK)

Decision tree (DT)

Directed acyclic graph (DAG)

European Centre for Disease Prevention and Control (ECDC)

Expectation-maximization (EM)

Explainable artificial intelligence (XAI)

Glasgow Coma Scale (GCS)

Gradient Boosting (GB)

Hospital-acquired bacteremia (HAB)

Hospital-acquired infection (HAI)

Hospital-acquired infections Database (HAIBA)

Hospital-acquired urinary tract infection (HA-UTI)

Incremental cost-effectiveness ratio (ICER)

Infection prevention and control (IPC)

International Classification of Diseases, Eighth- and Tenth Revision (ICD-8 and -10)

Local interpretable model-agnostic explanation (LIME)

Logistic regression (LR)

Machine learning (ML)

Neural networks (NN)

Numbers needed to harm (NNH)

Peter-and-Clark (PC) algorithm

Random forest (RF)

SHapley Additive exPlanation (SHAP)

Standard deviation (std)

Support vector machine (SVM)

Surgical site infections (SSI)

Tree-augmented naïve (TAN)

Urinary tract infections (UTI)

World Health Organization (WHO)

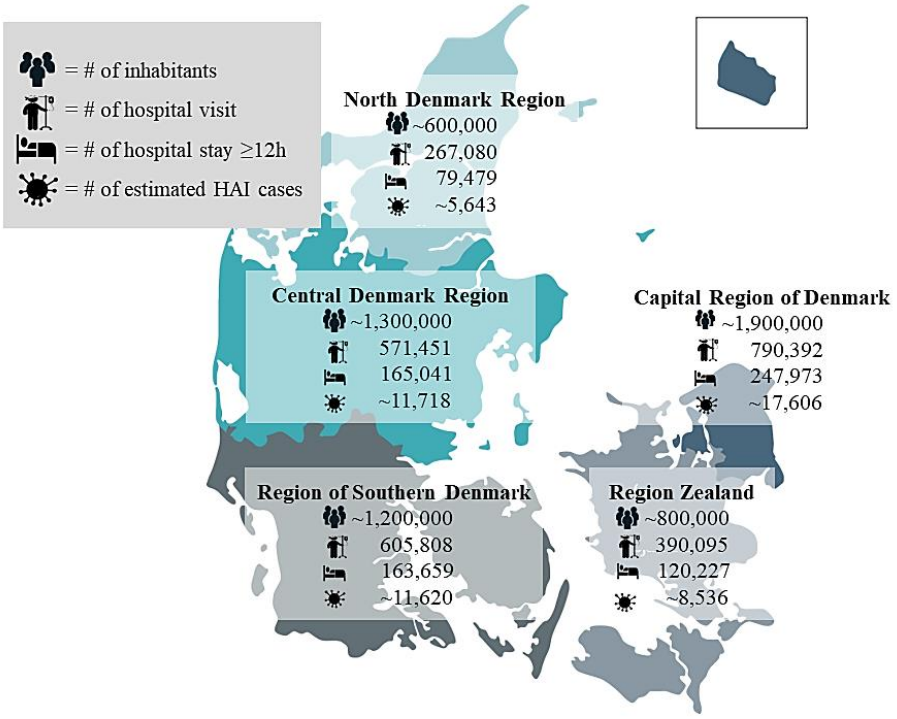
CHAPTER 1. BACKGROUND

The chapter introduces the theoretical background of this dissertation. First, an introduction to the domain of hospital-acquired infections (HAI) and approaches for management and control are presented, followed by outlining risk factors and recent measures for improved outcomes. Secondly, the perceptions of machine learning (ML) in the management and control of HAI are elucidated, including defining clinical explainable ML and the perspectives of responsible ML. Finally, the chapter outlines the significant challenges of antimicrobial resistance (AMR), emphasizing the need for innovative strategies in the future battle against preventable HAIs. This leads to the aim of this dissertation.

1.1. HOSPITAL-ACQUIRED INFECTIONS

HAI can be defined as the subset of infections acquired in the hospital that become evident 48h posterior to admission or shortly after hospital discharge [1–3]. They pose a major burden to healthcare providers worldwide [4–7] and are associated with high mortality rates, patient discomfort and disabilities, extended hospital stays, and high attributable costs [3–10]. The European Centre for Disease Prevention and Control (ECDC) estimates an average prevalence of 7.1 % of HAI, resulting from ~4,544,100 yearly episodes of these adverse instances across all European countries [4,5]. In Denmark, with a population of ~5,900,000 inhabitants [11], the prevalence of HAI varied between 6.5% and 9.2% from 2009 to 2014 [8], reaching an estimated annual ~60,000 episodes of HAI across Danish hospitals [12,13]. Figure 1 presents a map of Denmark with the distribution of inhabitants over the five regions, the yearly average of how many patients with a hospital visit, and how many of those with a stay ≥12h in each region (between 2009-2018) [14]. Figure 1 also presents an estimate of how many HAI cases that may be expected in a year, calculated by 7.1 % [4,5] of the average hospital stay ≥12h in each region:

Figure 1. A map of Denmark with the distribution of inhabitants over the five regions, the yearly average of how many patients with a hospital visit, how many of those with a stay ≥12h in each region, and an estimate of how many hospital-acquired infections cases that may be expected a year.[1]



Although it may seem contradictory, hospitals can sometimes be the origin of adverse health outcomes [6]. This may be due to several factors contributing to an elevated risk of infection for hospitalized patients, including reduced immunity due to illness, exposure to a range of medical procedures, and invasive techniques, which can provide an entry point for pathogens [2,6].

While the notion of *hospital-acquired* is used interchangeably with *nosocomial-*, *healthcare-associated-*, *or simply hospital infection* [4,8], *community-acquired infections (CAI)* are used to describe the remaining subset of infections [15]. The distinction between HAI and CAI may guide initial awareness of the origins of the infections for preventive purposes and, if necessary, the appropriate therapy [16,17]. Notably, at least ~20 % of all HAI may have been avoided, e.g., from proper

[1] Map modified from: https://rn.dk/genveje/fakta-om-nordjylland/regioner-i-danmark, seen April 2023). Hospital-acquired infection (HAI). Modified figure with numbers are self-made.

preventive strategies, significantly decreasing the burden of these adverse outcomes and enhancing patient safety [8,18]. In a 2015 systematic review conducted by Zingg et al. [5], ten key components were identified to mitigate HAI. World Health Organization (WHO) later refined these components in their eight core components for infection prevention and control (IPC) guidelines [19–22], which also applies to HAI. These components are grouped into four synergistic clusters in this dissertation as follows:

1. *Secure a collaborative and fair hospital environment:*
   The first cluster emphasizes the importance of fostering a hospital organization encompassing diverse expertise, such as nursing staff, infection control-trained physicians, laboratory personnel, and data management staff [5,19–23]. A fair distribution of ward occupancy should also be considered, as imbalances may result in high workloads among essential experts. Insufficient adherence to, e.g., hygienic measures can occur more frequently when understaffed individuals are burdened with excessive work hours [5,19–23]. A positive organizational culture also promotes a collaborative and fair environment, encouraging transparency and effective communication between experts and management levels [5,20,21]. Inconsistencies in, e.g., written and verbal communication may lead to negative consequences for patient safety [5].

2. *Ensure the physical settings of the hospital:*
   The second cluster focuses on facilitating access to necessary materials, such as hand sanitizers and alcohol swaps, and optimizing ergonomics within the hospital setting [5,19–23]. This includes providing customized insertion kits for central venous catheters and equipping carts with suitable materials [5]. Easing the accessibility of necessary materials promotes best-practices adherence [5,19–23].

3. *Educate and rely on best practices:*
   The third cluster emphasizes the measures of reaching best practices [5,19–23], e.g., the implementation of guidelines for proper hygienic measures [24] and antibiotic therapy [5,25–27]. While guidelines may not necessarily lead to behavioral changes, they may be implemented at the educational level to align ideas of therapies and avoid harmful contributions to, e.g., AMR [5]. In addition to providing guidelines, team- and task-oriented education and training for staff, including hands-on simulations for procedures, e.g., training in catheter replacements, may reduce errors that result in adverse outcomes [5,19–23]. Standardizing audit procedures are a core recommendation [5,19–23], where peers agree on, e.g., guidelines, education, and specific care bundles for catheter insertion for optimal outcomes [5].

4. *Welcome assistance:*
   The fourth cluster advocates utilizing multifaceted tools and techniques, e.g., by welcoming surveillance and feedback mechanisms to monitor the frequency of

occurrences and take appropriate measures for IPR [5,8,28]. Multimodal and interdisciplinary approaches to IPR [5,19–23], e.g., to utilize various data modalities from a diverse management and control strategy [5]. Finally, engaging with 'champions' of various processes, such as management experts for optimal leadership, developers of new technologies, and various implementation specialists, may result in better strategic planning for avoiding HAI [5].

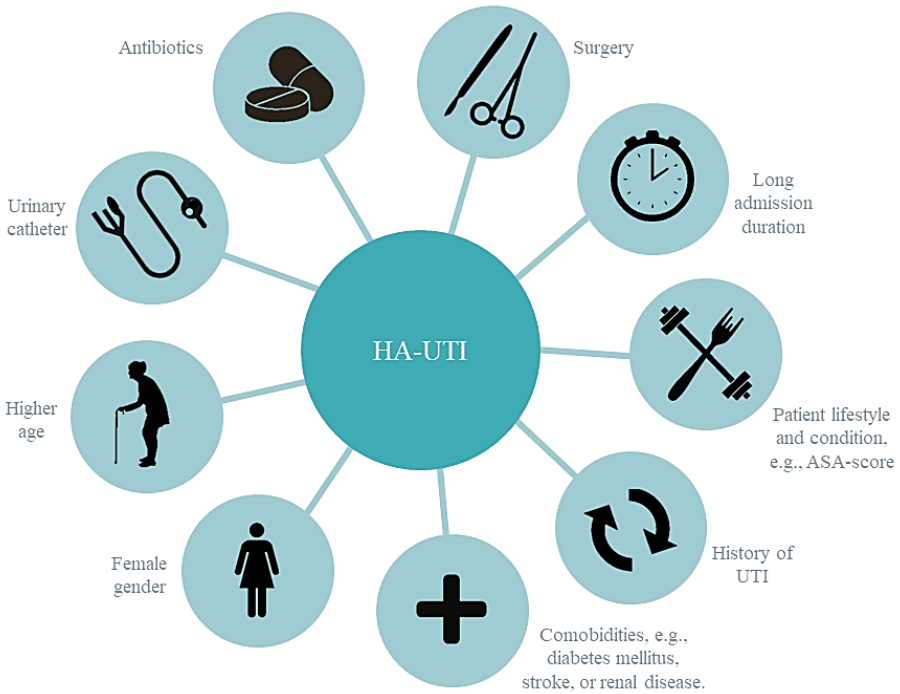
While HAI encompasses a comprehensive list of infection types [3], this dissertation will focus on preventing hospital-acquired urinary tract infection (HA-UTI), the most common type of HAI [4,29,30], and hospital-acquired bacteremia (HAB), which is associated with high mortality rates [31,32].

1.1.1. RISK FACTORS FOR HA-UTI AND HAB

HA-UTIs account for ~40% of all HAIs [4,29,30] and are associated with an attributable mortality rate of 9-13% [33,34]. Conversely, HABs have a reported 30-day mortality rate of 25–28 % [8,35] but are less common, with an incidence of 7.4 per 10,000 hospital patient days [36]. HA-UTI and HAB are further associated with extended hospital stays and heightened healthcare costs [4,37–39]. HAB may even originate secondary to an HA-UTI [34]. To enhance understanding and facilitate the prevention of HA-UTI and HAB, a wealth of research has been devoted to pinpointing risk factors and discerning patterns contributing to their incidence [30,40–42].

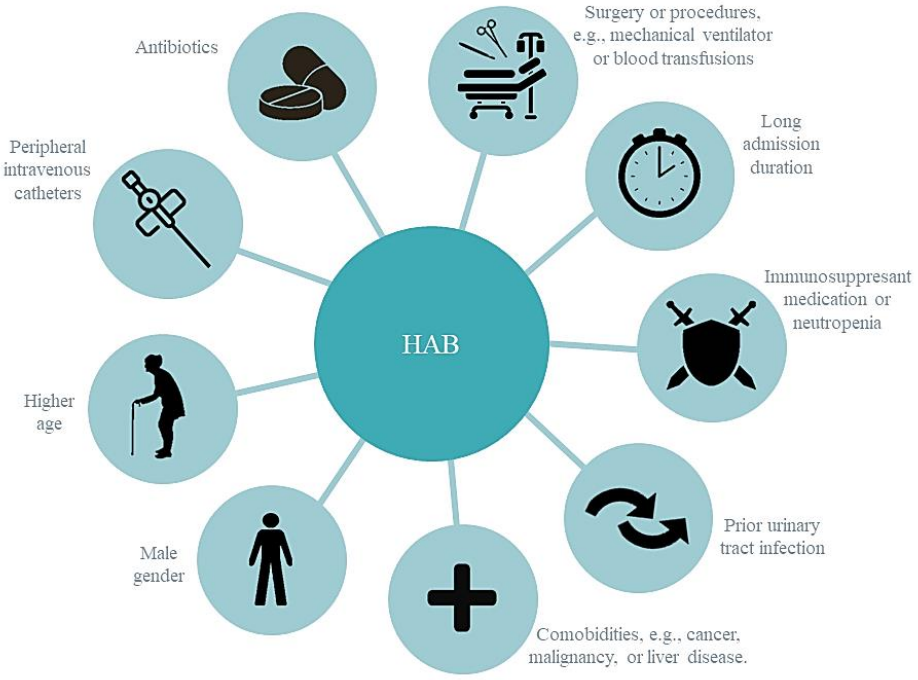
The utilization of urinary catheters has been identified as the most significant risk factor for hospital-acquired urinary tract infection (HA-UTI) development [41,43,44]. A 2012 review conducted by King et al. [41] suggested that a substantial 79.3% of all HA-UTIs might have been averted if urinary catheters had not been employed. Additionally, patients of advanced age [30,41,44], female sex [41,44], and those with comorbidities, such as diabetes mellitus, heart diseases, or neurological disorders, are also subject to a higher risk of HA-UTI [41,43,44]. A history of urinary tract infection [41,44] and prolonged hospital admissions [41] may also face an elevated risk of HA-UTI [30,41,43]. A higher American Society of Anesthesiologists (ASA) score, which results from lifestyle and patient condition, is also associated with an enhanced risk of HA-UTI [41,44]. The use of antibiotics and surgery are associated with a decreased risk [41]. Figure 3 illustrates the multifaceted factors that may affect the risk of HA-UTI:

Figure 2. Risk factors for hospital-acquired urinary tract infection.



For HAB, the usage of intravenous catheters is a significant risk factor, where up to 14 % of all HAB cases are associated with using a peripheral veinous catheter (PVC) [45,46]. Male gender [26], advanced age [26,47], and vulnerable patients with immunosuppressant medication or neutropenia [34], having specific comorbidities [26,47], or a urinary tract infection (UTI) [34] are also considered at risk of HAB. Surgery or procedures [26,34], such as blood transfusions or mechanical ventilation, are also at enhanced risk of HAB [34], whereas antibiotics decrease the risk of HAB [34]. Figure 4 illustrates the multifaceted factors that may affect the risk HAB:

Figure 3. The risk factors for hospital-acquired bacteremia.

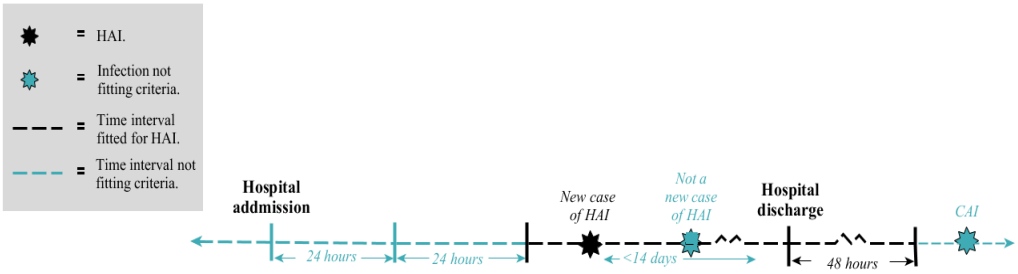


1.1.2. SURVEILLANCE SYSTEMS

The healthcare system has undergone a significant digital transformation in recent decades [48], opening up new opportunities for the strategy to counter HAI [49]. One remarkable example of such an opportunity is the development of comprehensive surveillance systems to monitor incidence rates and enable benchmarking between hospitals [8,28]. In a review by Sips et al. [28] on surveillance systems for HAI, it is signified that assistance by conveying such an overview would not be any better than the case definition of what attempted to be tracked. Relying on stand-alone administrative data for case definitions of HAI will not be appropriate, whereas strategies combining multiple high-quality administrative and clinical data sources revealed more consistent results [28]. In Denmark, for example, a recently implemented national surveillance system, called Healthcare-Associated Infections

Database (HAIBA) [8], registers an HAI if certain administrative-, microbiological- and clinical conditions are met, previously described by Gubbels [8], within >48h posterior to admission and <48h after discharge, with no HAI occurring 14 days prior to a new case of HAI (Figure 4).

Figure 4. Timeline with stamps for an infection to be included as a hospital-acquired infection in HAIBA[2].



The list of monitored types of HAI in HAIBA includes HA-UTI, HAB, *Clostridioides* (former *Clostridium*) *difficile* infections (CDI), and prosthetic joint infections [8,10,50], making their case definitions and registrations in the database suited for the purposes of this dissertation. In this context, Gubbels [8] writes on future perspectives in their work on HAIBA:

> *(1) "The current case definitions and output models form a good basis to build further upon, but have by no means reached the end of their potential."* [8]

1.2. MACHINE LEARNING AND HOSPITAL-ACQUIRED INFECTIONS

The potential benefits of computational mathematical models in utilizing healthcare data are considered extensive [51], holding the promises of a more efficient and effective healthcare system with improved patient outcomes in general [48,51,52]. Traditionally, various expressions are used to describe these data-driven and computational approaches in a larger inter-connected ecosystem, including *artificial intelligence (AI)*, *machine learning (ML)*, *pattern recognition*, *deep learning*, *data mining*, *Big Data*, and the list goes on [52,53]. Moreover, by integrating these models into the healthcare domain, solutions may relate to the notions of *biomedical*- or *healthcare informatics*, e.g., in the sense of *clinical decision support systems* (CDSS) [48,51] or *predictive analytics monitoring* [52]. In the book by Dr. Eric Topol in 2019 called '*Deep Medicine: how artificial intelligence can make healthcare human again'*,

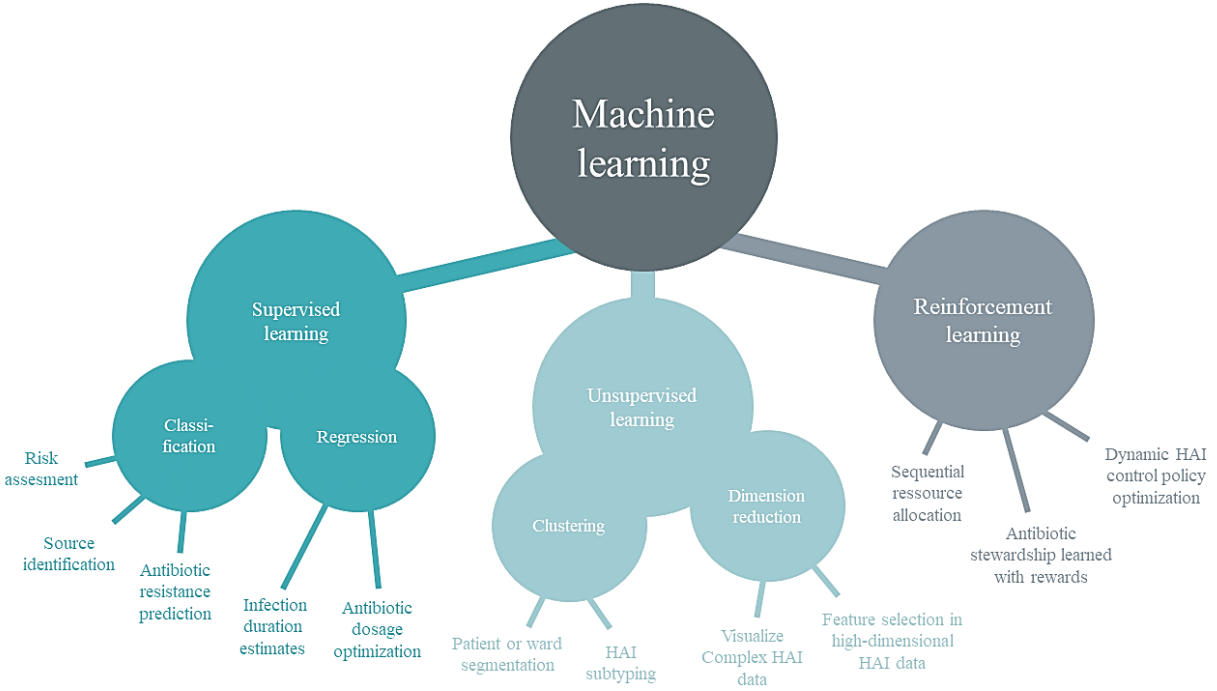
[2] Hospital-acquired infection (HAI). Community-acquired infection (CAI).

the term *deep medicine* was also coined as a hybrid expression between deep learning and medicine [54], which may also invite a further future expansion in the palette of terms. Broadly speaking, the goal for this inter-connected ecosystem of computational mathematical models is:

(2) "To identify patterns in data, and then learn from the identified." [48]

With the increased availability of vast amounts of healthcare data [51,55], and the perspectives of identifying patterns and then learning from the identified, new opportunities may also arise for timely and targeted management and control of HAI [49,56,57]. Dependent on the type of task that needs solving, ML is traditionally categorized into three types of learning: supervised-, unsupervised-, and reinforcement learning [53,56,57]. Supervised learning refers to ML models trained with a known labeled target (e.g., HA-UTI or HAB), whereas unsupervised learning refers to ML models trained without such a labeled target [53]. Reinforcement learning refers to ML models that learn to make decisions by interacting with the environment, e.g., from rewards [53]. Figure 5 visualizes ML and its categories of learning, sub-categories and examples of decision support perspectives of in control and management of HAI.

Figure 5. Machine learning in control and management of hospital-acquired infections, its categories of learning, and sub-categories with examples of decision support[3].



[3] Machine learning models can be categorized into supervised-, unsupervised- and reinforcement learning. Supervised learning are further categorized into classification or regression, whereto unsupervised learning are further categorized into clustering or dimension reduction. The different categories of learning may results in different perspectives of decision support for ML in the control and management of hospital-acquired infection (HAI).

Despite ML models' perspectives on the management and control of HAI, the cross-disciplinary role is remarkably unexplored and, as a result, not entirely understood [49]. Two systematic reviews for ML for infectious diseases by Peiffer-Smadja et al. [56] and Luz et al. [57] includes overlapping 60- and 52 studies, respectively, but from a broader scope of infection inclusion criteria. The predominant part relies on supervised learning (>90 %); the remaining on unsupervised learning, and only a small part on reinforcement learning. Most studies (>60 %) are about bacterial infections; the remaining are related to viral infections, tuberculosis, fungi, or others. The majority of studies are related to risk assessment or various kinds of therapy responses, while other studies are explicitly related to resistance, surgical site infections (SSI) and other postoperative infections, or specific to HAI [56,57].
Tang et al. [58] conducted a systematic review of ML specific to antibiotic resistance, including 25 studies on their content, but only one study particularly related to HAI. A systematic review specific to HAI by Scardoni et al. [49] includes 27 studies, one-third on HAI in general and the remaining on a specific type of HAI, including a majority of overlapping references to Peiffer-Smadja et al. [56] and Luz et al. [57] on SSI and sepsis. Quite remarkably, across all meta-analyses, future directions relate to exploring the fit and improving the understanding of adaption into the clinical routine [49,56–58]. This may, in particular, relate to addressing dimensions of explainability and interpretability to establish trust for clinicians in these new technologies [57]. If the clinicians are challenged in interpreting the given model contribution, the models remain unlikely to reach the daily clinical routines [49,56,57,59–65].

Studies specific to ML for HA-UTI and HAB are generally rare. Jeng et al. presented an approach to predict recurrent urinary tract infections (UTI) [66] using random forest (RF), decision tree (DT), logistic regression (LR), support vector machine (SVM), and neural network (NN), suggesting that RF reaches the best performance for their purpose. Also, Møller et al. [67] aimed at early prediction of HA-UTI within 48h of admission, comparing a selected number of ML approaches of what they called *easily interpretable* and having *complex interpretation* in early prediction of HA-UTI. Their recommended DT reached an area under the curve (AUC) of 0.709, whereto their best-performing NN reached an AUC of 0.77. Zhu et al.[68] explored ML in predicting poststroke UTI risk in immobilized patients from a pipeline of resampling techniques for their imbalanced dataset. They compare the performance of various ML models, reporting that RF reaches the best performance with an AUC of 0.82, also assessing the feature importance of their models, revealing that risk factors of pneumonia, glucocorticoid use, female sex, specific comorbidities, advanced age, prolonged hospitalization, and catheterization. Other studies on ML models and UTI have also been conducted, including Taylor et al.[69] comparing predictive performance between seven algorithms for urinary tract infections in the emergency department, suggesting XGboost as the preferred approach due to better performance. Yelin et al.[70] also demonstrated XGboost in their work, which is compared to LR, but to identify antibiotic resistance for UTI. Lastly, Park [71] developed predictive models specific to catheter-associated urinary tract infections, suggesting that DT may be a better predictive model than SVM and LR for their purpose, and that female gender and advanced age are the most significant risk factors.

For HAB, Garnica et al. [72] sought to enable timely and targeted preventive measures from ML developed over a relatively sparse dataset, which may reflect their reported AUCs above 0.90. Other studies have compared the ML model's performance in assessing the risk of central line-associated bacteremia [58,59], predicting bacteremia without distinguishing between community- and hospital-acquired cases [72–76], or whether the ML model could identify patients with bacteremia who met the systematic inflammatory response syndrome criteria [77]. However, despite posing a potential for ML models in timely and targeted prevention, no study except Garnica et al. [72] is specific to HAB.
Of notice, studies for HA-UTI and HAB rarely address dimensions of explainability or feasibility, leaving these critical perspectives [57] poorly understood.

1.2.1. CLINICAL EXPLAINABILITY

In recent years, various papers have been published that each tells a story of explainable ML [60,63,78–85], e.g., to promote trust [59–61], fairness [59,60,86], transparency [59,80], accountability [60,63,86,87], informativeness and functionality from robustness, reliability or usability [60,82]. However, to pursue explainability, a trade-off may have to be addressed with the model performance [78,86], setting the stage for a complex assessment of the model's fit into the clinical routine. Furthermore, while *explainability*, *interpretability*, *understandability*, and even *intelligibility* are often used synonymously, a universally accepted definition remains elusive, leaving the concept somewhat ill-defined [59,82,86]. This dissertation adopts the definition in the book by Shaban-Nejad et al. [80] from 2020 called *'Explainable AI in Healthcare'*, and refines it specifically for the medical environment with diverse expertise, hence defining *clinical explainability* in ML as:

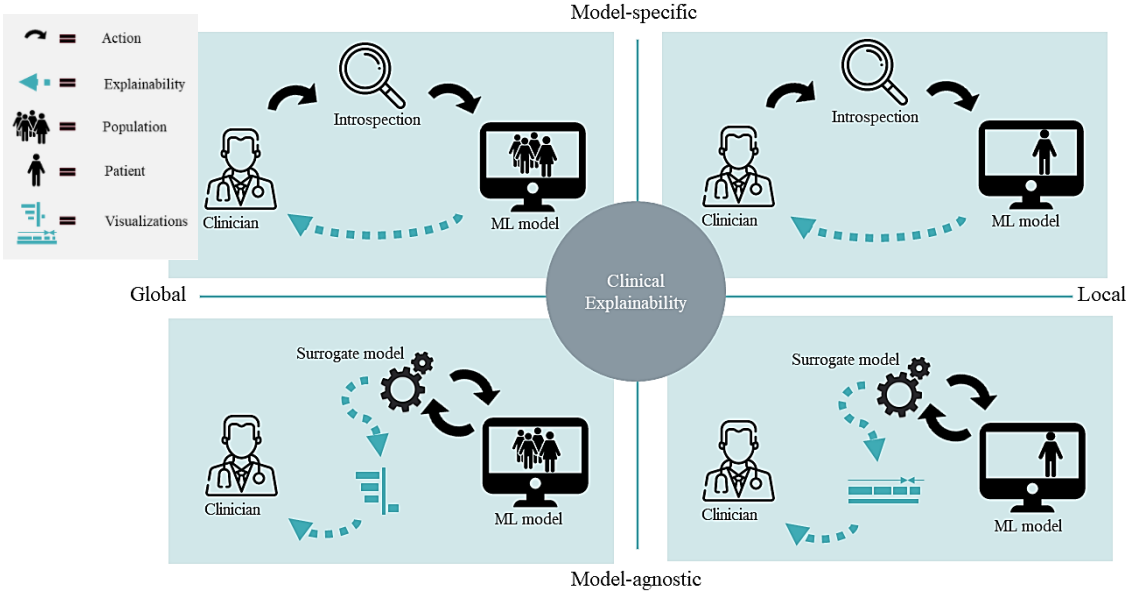
> *(3) "The capacity of a model to be comprehended by clinicians across corresponding hospital wards, either through introspection and/or a produced explanation."*

While definition (3) focuses on the heterogenous needs and experiences across the hospital organization's diverse areas of expertise, it also reflects the critical differentiation between model-specific and model-agnostic explainability [60,82,84], which can be further categorized into global- or local-level contexts, depending on whether they describe the ML model over a population or a single prediction [60,84,86].
Model-specific techniques involve analyzing the internal mechanics (introspection), which only applies to a subset of ML models classes, such as Bayesian network (BN), LR, and DT, where one can explore the inner mechanics of the model, e.g., by inspecting feature dependencies, model parameters, or the rules that define the splits in a tree [60,78,84]. This further enables a hands-on assessment of the ML model structures, e.g., to incorporate clinical expert knowledge (heterogeneous needs and experiences) in the model's design [63] and potentially address the critical dimensions

of causalities [59,60,85,86]. Of notice, Rudin et al. [83] describe these model classes as *inherently interpretable* and highlight their use over models that do not meet these qualities, hence being black-box, to avoid bad practices by not understanding the inner mechanics of the model, despite being provided with an explanation of their clinical reasoning.

Model-agnostic techniques evaluate a given ML model's predictions as ante-hoc to its given result (produced explanation) [60,84,88], which can be applied as an extension to any ML model class, including NN and ensemble methods. Model-agnostic techniques may be associated with assessments of feature importance [82], e.g., from SHapley Additive exPlanation (SHAP)-method [89] local interpretable model-agnostic explanation (LIME)-method [61], or as counterfactual explanations [65,90]. Assessing feature importance in model-agnostic explainability gives a sense of which feature has the highest impact on the model's predictions, whereas counterfactual explanations provide insight into what changes could lead to a different model prediction [65]. Of notice, model-agnostic explainability has the advantage of not sacrificing performance by relying on less complex, yet easier interpretable, ML models [59]. Figure 6 exemplifies the distinction between model-specific and model-agnostic explainability at local- and global levels.

Figure 6. Model-specific and model-agnostic explainability of a machine learning model at global and local levels[4].



[4] Inspection of, e.g., feature dependencies model parameters, the rules that define the splits are examples of global model-specific clinical explainability. Local model-specific clinical explainability may be supported by inspecting evidence in nodes, or decision path in for a single given patient admission. Global model-agnostic explainability may be supported from a visualization of how sensitive the ML model prediction are for the given feature values. Local model-agnostic clinical explainability may be supported by a visualization of how each feature has contributed to the ML model prediction for a single patient admission.

While clinical explainability in ML for the sensitive healthcare domain may be obvious [60], no studies have been conducted for such approaches in mitigating HA-UTI or HAB. Moreover, clinicians may lack confidence, experience, and the vital aspects of trust in the predictions of these models [64,91]. As a result, the deployment into the clinical routine remains challenging in various steps of developing a model [64]. The aspects of explainability should be responsibly considered throughout the entire lifecycle of the model, encapsulating the concept of responsible ML [64,92].

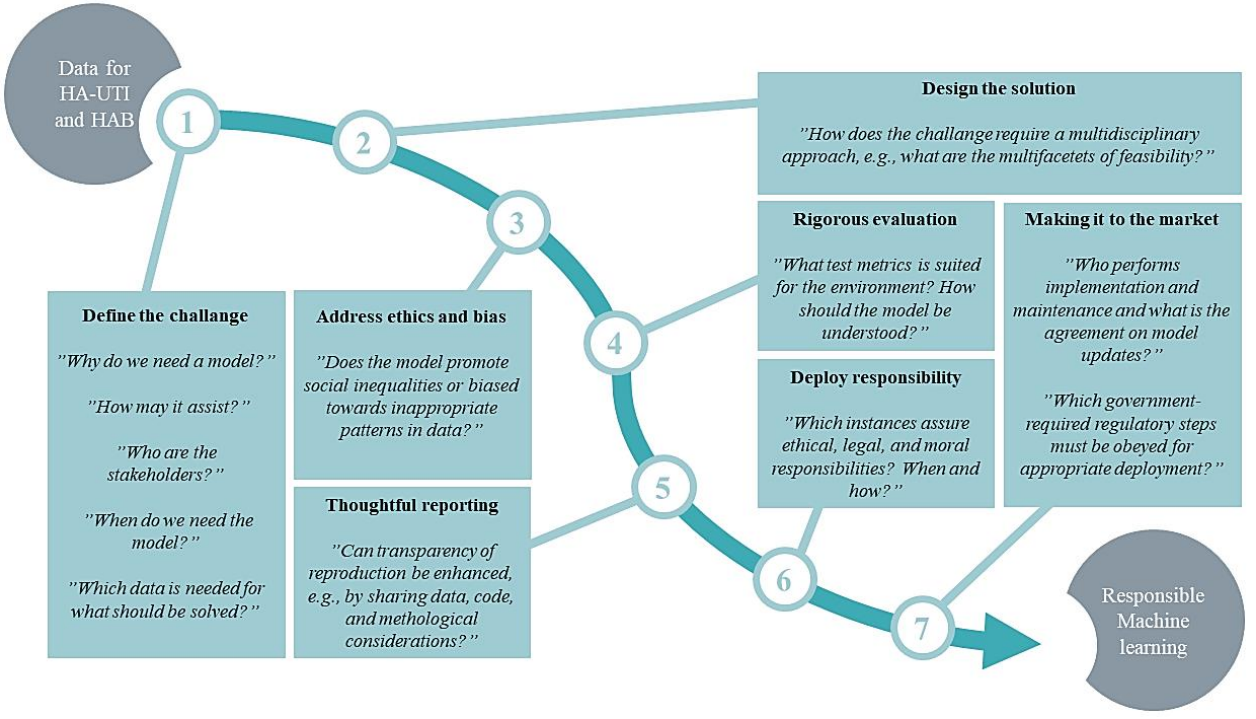
1.2.2. RESPONSIBLE MACHINE LEARNING

To promote chances of adaption into the clinical routine, the realm of responsible ML [64,92] has been suggested as the backbone for a process from the cradle to the grave. In this scenario, Wiens et al. [64] define a seven-step process in the paper *'Do no harm: A roadmap for responsible machine learning for healthcare'*, which also applies to responsible ML models for HA-UTI and HAB.

The first step is establishing common ground and understanding the environment and the ideation. While Ribeiro et al. [61] and Gilpin et al. [93] discussed the *why*-question related to explaining a model prediction, the models fit into the clinical routine, hence *why do we need the model* in the first place, should be answered in addition to *how*, *who*, *when*, *where* and *what* [80,85,94]. The *who*-question invites an expectation of a multidisciplinary approach with several stakeholders, e.g., clinical staff of diverse expertise, ML developers, patients, hospital administration, health economists, implementation experts, etc. [64]. This is also the second step of the process of responsible ML [64]. Of note, in the review by Scardoni et al. [49] on ML for HAI, 74 % was the product of multidisciplinary approaches between, e.g., clinical and IT researchers, but without presenting model designs, ideas of explainability, or dimensions of feasibility. The third step is to inspect the data, e.g., in evaluating whether the model may conflict with ethical standards or be subject to bias [80]. The fourth step invites rigorous evaluation, which includes assessing how to handle, e.g., imbalanced data with missingness, the robustness of case definition (e.g., relying on HAIBA), validation approach, and how data may be split while avoiding leakage that may lead to overfitting [64]. This step may resolve from traditional best practices in ML model development [53].

The fifth step is carefully deciding how the ML model output may be available to the clinician, e.g., choosing performance metric and explainability approach [64]. McCoy et al. [95] further emphasize the need to educate staff to comprehend ML models' outcomes. The sixth step suggests deploying responsibilities, which also bridges a research environment and implementation into the real-world clinical routine. This also includes that the multidisciplinary team members identified in step one are responsible for ensuring ethical, moral, and legal pipelines so that an ML model fits the environment [64]. The seventh step connects the model to the society's governmental regolatories and suggests guidelines for maintaining the model and correcting possible errors, or how and when a model may be retrained on updated data [64]. Figure 7 presents the process of responsible ML for the aims of this dissertation:

Figure 7. The process of responsible machine learning for healthcare, including questions that may arise in the seven-step approach[5]



[5] Hospital-acquired urinary tract infection (HA-UTI). Hospital-acquired bacteremia (HAB).

1.3. GRAND CHALLENGES

While HAIs are associated with high mortality rates, patient discomfort and disabilities, extended hospital stays, and high attributable costs [3–10], they also pose severe challenges regarding AMR, an issue that is anticipated to become increasingly critical within the coming years [96]. In this context, ECDC sounded an early warning in their report in 2009 called *'The bacterial challenge, time to react'* [33], wherein they emphasized the urgent need to address evolving AMR patterns in Europe. The urgency of this situation was further underlined in the 2014 report by O'Niell called *'Antimicrobial Resistance: Tackling a Crisis for the Health and Wealth of Nations'* [97] estimated that by the year 2050, 10 million people will die due to bacterial infection and AMR. Notably, this figure exceeds the projected number of deaths from cancer by two million [97,98]. Moreover, the impact of these challenges is expected to extend beyond the medical sector, spilling over into, e.g., social- and economic domains [96,99].
If the premise of at least ~20 % of all HAIs may be prevented [8,18] – a prevention not currently being achieved – it may be obvious that innovative tools are needed that enhance the capability to mitigate HAI. This sets the stage for ML in the management and control of HAI [48,49,56,57]. However, despite the promising performances of the proposed ML models in HAI, they still struggle to find the daily clinical routine.

Figure 8. AI-generated illustration by DALL.E 2 from the command *A doctor relying on screens with computer-generated risk scores for a patient in bed, digital art*.

This dissertation explores the development of clinical explainable ML models for enabling timely preventive measures of HA-UTI, the most common type of HAI

[4,29,30], and the feasibility of ML models in HAB prevention, the HAI associated with high mortality rates [31,32]. The ML models are developed from data curated from the North Denmark Region. At the heart of this dissertation lies the perception that if clinicians do not understand the ML model, it remains unlikely that they will trust them, resulting in the unlikeliness of reaching the daily clinical routine.

CHAPTER 2. AIMS

In this dissertation, it is hypothesized that (1) ML models can be developed that reach high performances and still support model-specific clinical explainability, (2) model-agnostic explainability can be supported for different ML models with different performances, and (3) ML models can be developed for guided PVC replacement strategy, whereto numbers needed to harm (NNH), and cost-effectiveness address vital aspects of gaining insight on feasibility.

Study I:

The aim of study I is to compare the performance of a single ML model type, being BN complemented for model-specific explainability, trained from different methods of structure learning for predicting the risk within 24h of admission of HA-UTI. Furthermore, to simplify the BN models with fewer features, a feature selection is performed by combining clinical expert knowledge and machine suggestions, which is compared to BN models developed over a full dataset of 51 features. Model-specific explainability is addressed by discussing the ability to find dependencies in data that reflect a domain.

Study II:

Study II aims to compare different ML models' performance in predicting the risk within 24h of admission of HA-UTI and how model-agnostic explainability may be applied to explain the model predictions. The impact on the performance from different methods for feature selection prior to model development is also evaluated. The SHAP method is used to explore both global- and local levels of model-agnostic explainability.

Study III:

Study III aims to compare different ML models' performance and feasibility in predicting the risk within 24h of admission. Patients at risk of HAB may benefit from the routine replacement of PVC, which is suggested to prevent more HAB. In contrast, patients not at risk of HAB may be subject to clinical indicated replacement strategies, which are suggested to be more time-saving. Feasibility is assessed in the associated NNH and the incremental cost-effectiveness ratio (ICER) of implementing routine replacement of PVC every 96h when guided by ML risk prediction.

CHAPTER 3. MATERIALS AND METHODS

This chapter presents the materials and methods used in all three papers. First, an introduction to study designs, data sources and participants, definitions of HA-UTI and HAB outcomes, discretization, and the model design for enhanced simplicity are presented. These topics set the foundation for all three papers. This is followed by an introduction to the materials and methods in Study I, where an ML model with characteristics of enabling model-specific clinical explainability in risk prediction is explored. Study II presents the comparison of different ML models and how clinical explainability can be supported in model-agnostic clinical explainability. In study III, different ML models were compared to risk stratify HAB and guide the replacement strategy for PVC, which includes exploring the feasibility of such a purpose.

3.1. STUDY DESIGNS

The retrospective studies conducted in this dissertation are in a collaboration between the Centre for Clinical Research at the North Denmark Regional Hospital, the Business Intelligence and Analysis Unit in the North Denmark Region, and the Department of Computer Science at the University of Aalborg in Denmark. For study III, a collaboration is also established with the Danish Center for Health Services Research at the University of Aalborg and the Department of Clinical Microbiology, Aalborg University Hospital.

Danish legislation requires no approval from an ethics committee or participant consent for registry-based studies.

3.1. DEFINITION OF HA-UTI AND HAB OUTCOMES

The outcome of HA-UTI in Study I and Study II, and HAB in Study III, are binary features in the datasets, available through a collaboration with HAIBA [8,10]. The HA-UTI- and HAB definitions in HAIBA are based on a combination of time stamps and blood culture findings previously described in detail by Gubbels et al. [8,50]. Timestamps for HA-UTI and HAB are presented in Figure 4.

3.2. SOURCE OF DATA AND PARTICIPANTS

Adults between 18 and 100 years of age with a length of stay between 0.1 and 365 day(s) at North Denmark Regional Hospital or Aalborg University Hospital in Denmark from 1 January 2017 to 31 December 2018 are included in the three studies. Data are linked using the unique civil personal registration (CPR) number issued to all Danish citizens and the timestamps for the hospital admission. A patient might be included more than once in our dataset if the patient had experienced more admissions at different time points during the two-year study period. Statistical Analysis Software Enterprise Guide 8.1 is used for data management.

In Denmark, the medical condition leading to hospital admission is registered as an A-diagnosis, and if another condition occurs, a B-diagnosis is added for each additional condition. Since 1994, diagnoses have been registered in accordance with the International Classification of Diseases, Tenth Revision (ICD-10). Diagnoses registered before 1994 are registered in accordance with -eighth revision (ICD-8). A level of three digits in ICD-10 for A-diagnosis is used to construct an *admission cause*-feature. Because *External causes of morbidity and mortality* and *Certain conditions originating in the perinatal period* are not used for primary cause of admission and patients in the dataset are > 18 years of age, respectively, these are excluded from the dataset. The ICD-10-codes used for admission causes are presented in Table 1.

Table 1. The cause of admissions and the associated International Classification of Diseases, Tenth Revision, codes[6]

#	Cause of admission	ICD-10 codes
1	Certain infectious and parasitic diseases	A00 – A99, B00 – B99
2	Neoplasms	C00 – C99, D00 – D49
3	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	D50 – D99
4	Endocrine, nutritional and metabolic diseases	E00 – E99
5	Mental and behavioral disorders	F00 – F99
6	Diseases of the nervous system	G00 –G99
7	Diseases of the eye and adnexa	H00 – H59
8	Diseases of the ear and mastoid process	H60 – H95
9	Diseases of the circulatory system	I00 – I99
10	Diseases of the respiratory system	J00 – J99
11	Diseases of the digestive system	K00 – K93
12	Diseases of the skin and subcutaneous tissue	L00 – L99
13	Diseases of the musculoskeletal system and connective tissue	M00 – M99
14	Diseases of the genitourinary system	N00 – N99
15	Pregnancy, childbirth and the puerperium	O00 – O99
16	Certain conditions originating in the perinatal period	P00 – P96

[6] The International Classification of Diseases, Tenth Revision (ICD-10)

17	Congenital malformations, deformations, and chromosomal abnormalities	Q00 – Q99
18	Symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified	R00 – R99
19	Injury, poisoning, and certain other consequences of external causes	S00 – S99, T00 – T98
20	External causes of morbidity and mortality	X60 – X98, Y00 – Y09
21	Factors influencing health status and contact with health services	Z00 – Z99

The selected comorbidities are inspired by the Charlson Comorbidities Index (CCI), whereto the associated ICD-10 for CCI [100–102] and the ICD-8 codes for CCI [102] are used to define the included comorbidities in the dataset. In addition, a history of UTI is also included for Study I and Study II, with the associated ICD-10- [8,103] and ICD-8 codes [103]. The *human immunodeficiency virus* (HIV) is not included due to its rare occurrences. The comorbidities and their associated ICD-10 and ICD-8 codes are presented in Table 2.

Table 2. The Charlson comorbidity index and the associated International Classification of Diseases, Eighth- and Tenth Revision, codes[7]

#	Features	ICD-10 codes	ICD-8 codes
1	Acute myocardial infarction	I21 – I23, I25	410
2	Congestive heart failure	I50, I11, I13	042709 – 42711, 42719, 42899, 78249
3	Peripheral vascular disease	R02, Z95, I70, I71 – I79	440 – 445
4	Cerebral vascular accident	I60 –I67, G45, G46	430 – 438
5	Dementia	F00 –F03, F05, G30	29009 – 29019, 29309
6	Pulmonary disease	J00 – J86, J90 – J99	490 – 493, 515 – 518
7	Connective tissue disorder	M05 – M14, M30 – M36, D86	13599, 4446, 712, 716, 734
8	Peptic ulcer	K22, K25, K26, K27, K28	53091, 53098, 531, 532, 533, 534
9	Liver disease	B17, B18, K71, K73, K74, K75, K77	571, 57301, 57304
10	Diabetes	E10, E100, E101, E109, E109A, E11, E110, E111, E119, E13, E130, E131, E135 – E139, E14, E140 – E149	24900, 24906, 24907, 24909, 25000, 25006, 25007, 25009
11	Diabetes complications	E102 – E105, E105A, E105B, E105C, E105D, E106 – E108, E112 – E118, E132, E133, E134, E142 – E148	24901 – 24905, 24908, 25001 – 25005, 25008
12	Paraplegia	DG04, DG80, DG81, DG82, DG83	344
13	Renal disease	N01, N02, N03, N04, N05, N06, N07, N08	403, 404, 580 – 584, 59009, 59319, 75310 – 75318, 792
14	Cancer	C00 – C76, C86	140 – 194, 19248
15	Metastatic cancer	C77 – C80	195 – 199
16	Leukemia	C91 – C95	204 – 207
17	Lymphoma	C81 – C85, C88, C90, C96	200 – 203, 27559
18	Severe liver disease	B15, B16, B19, K70, K72, K76, I85	07000, 07002, 07004, 07006, 07008, 45600 –45609, 57300, 07983
19	HIV	B20 – B24	-
20	History of UTI	N30, N390, N390, O234, P393, O233, O239, O862, T814U	580, 590, 59500, 59501

[7] The International Classification of Diseases, Tenth Revision (ICD-10). The International Classification of Diseases, eighth Revision (ICD-8)

3.2.1. DISCRETIZATION

Discretization of features is performed in Study I and Study II, whereto age, BMI, vital parameters, and laboratory results are included as continuous features in Study III. Only the worst measure is included in the discretized dataset. However, determining the *worst* measurement may depend on the degree of abnormality for certain vital parameters and laboratory results. This introduces complexity in defining discrete levels – for example, temperature measurement is two-sided, implying the *worst* measurement could be either abnormally high or low. Consequently, the *worst* measurement is dictated by the level of abnormality (e.g., a discrete level of '< 35°' is deemed worse than '38° - 38.9°'). If two measurements are performed within the same day – one at discrete level 3 and the other at level 4 – only the measurement from level 4 is included because it is deemed *worst*. In this scenario, determining appropriate discrete levels in the dataset presents challenges, such as establishing the proper quantity of discrete steps and defining thresholds between them. This dissertation employs expert-based knowledge to address these issues, incorporating recent literature findings that include pre-existing scoring systems used in intensive care units [104]. The discrete levels of the features are presented in Table 4.

3.1. MODEL DESIGNS

In each study, the ML models aim to facilitate patient risk stratification within the first 24 hours of admission. This approach serves as an alert mechanism to identify patients who could benefit from preemptive measures instituted early in their hospital stay. It serves as a strategic tool for enhanced awareness. Figure 10 presents an illustration of the model designs.

Table 3. Descitization of the predictive features [8]

	Feature	Discrete level 0	Discrete level 1	Discrete level 2	Discrete level 3	Discrete level 4	Discrete level 5	Discrete level 6	unit
	Admission detail								
1	Triage	Blue	Green	Yellow	Orange	Red	-	-	-
2	Season	Winter	Fall	Spring	Summer	-	-	-	-
	Health socio-demographics								
3	Age	18 – 29	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	> 80	Years
4	Gender	Female	Male	-	-	-	-	-	-
5	Body Mass Index	< 18.49	18.5 – 24.9	25. – 29.9	30 – 34.9	35. – 39.9	> 40	-	kg/m^2
6	Smoking	No	Ex-smoker	Yes	-	-	-	-	-
7	Alcoholic	No	Yes	-	-	-	-	-	> 14w, >21m[9]
8	Exercise	No	Yes	-	-	-	-	-	> 30min/d[10]
	Comorbidities								
9	Number of comorbidities	0	1	2	> 2	-	-	-	
10	Acute myocard. infarction	No	Yes	-	-	-	-	-	-
11	Congestive heart failure	No	Yes	-	-	-	-	-	-
12	Peripheral vascular disease	No	Yes	-	-	-	-	-	-
13	Cerebral vascular accident	No	Yes	-	-	-	-	-	-
14	Dementia	No	Yes	-	-	-	-	-	-
15	Pulmonary disease	No	Yes	-	-	-	-	-	-
16	Connective tissue disorder	No	Yes	-	-	-	-	-	-
17	Peptic ulcer	No	Yes	-	-	-	-	-	-
18	Liver disease	No	Yes	-	-	-	-	-	-
19	Diabetes	No	Yes	-	-	-	-	-	-

[8] The admission cause feature is presented in Table 1.

[9] The limit of being alcoholic is set to >14 drinks a week for women and > 21 drinks a week for men.

[10] The limit for exercising sufficiently is set to > 30 minutes per day.

20	Diabetes complications	No	Yes	-	-	-	-	-	-
21	Paraplegia	No	Yes	-	-	-	-	-	-
22	Renal disease	No	Yes	-	-	-	-	-	-
23	Cancer	No	Yes	-	-	-	-	-	-
24	Metastatic cancer	No	Yes	-	-	-	-	-	-
25	Leukemia	No	Yes	-	-	-	-	-	-
26	Lymphoma	No	Yes	-	-	-	-	-	-
27	Severe liver disease	No	Yes	-	-	-	-	-	-
28	History of UTI	No	Yes	-	-	-	-	-	-
	Vital parameters								
29	Temperature	36 – 37.9	35. – 35.9	38. – 38.9	39 – 39.9	> 40.	< 35.	-	^{O} Celsius
30	Glasgow Coma scale	15	14	9 – 13	3 – 8	-	-	-	-
31	Respiratory rate	12 – 20	8 – 11	21 – 24	25 – 29	30 – 35	> 36	< 8	Breaths pm
32	Pulse	50 – 89	90 – 109	40 – 49	110 – 129	> 130	< 40	-	Beats pm
33	Systolic blood pressure	110 – 219	100 – 110	90 – 99	> 220	< 90	-	-	mmHg
34	Diastolic blood pressure	60 – 89	90 – 99	100 – 119	> 120	< 60	-	-	mmHg
35	Oxygen saturation	> 96	94 – 95	92 – 93	80 – 91	< 80	-	-	%
	Laboratory results								
36	B-thrombocytes	145 – 389.9	60 – 139.9	390 – 999.9	20 – 59.9	< 19.9	> 1000	-	x 10^{9}/L
37	B-erythrocytes	3.9 – 5.69	5.7 – 7.49	> 7.5	2.5 – 3.89	< 2.49	-	-	x 10^{12}/L
38	P- partial pressure of oxygen	> 9.33	8. – 9.32	7.33 – 7.99	< 7.32	-	-	-	kPa
39	P- C-reactive protein	< 7.99	8. – 19.99	20. – 39.9	40. – 199.9	> 200.	-	-	Mg/L
40	B-hemoglobin	7.3 – 10.49	4. – 7.29	10.5 – 12.49	> 12.5	< 3.99	-	-	mmol/L
41	P-albumin	34 – 47.99	> 48.	16 – 33.99	> 16	-	-	-	g/L
42	P-creatinine	45 – 104.99	20 – 44.99	105 – 119.99	< 19.99	> 120	-	-	µmol/L
43	P-bilirubin	5 – 24.99	< 4.99	25 – 59.99	60 – 79.99	> 80	-	-	µmol/L
44	P-glucose	4 – 7.99	8 – 11.99	12 – 19.99	20 – 32.99	2 – 3.99	> 33	< 2	mmol/L
45	P-pH	7.33 – 7.49	7.5 – 7.59	7.25 – 7.32	7.6 – 7.69	7.15 – 7.24	< 7.14	> 7.7	-
46	P-lactate	0.5 – 2.49	< 0.49	2.5 – 8.99	> 9.	-	-	-	mmol/L
47	B-leucocytes	3.5 – 8.79	8.8 – 19.99	20 – 49.99	0.5 – 3.49	> 50	< 0.49	-	x 10^{9}/L
48	B-neutrofilocytes	2 – 8.79	0.5 – 1.99	8.8 – 19.99	20 – 49.99	> 50	< 0.49	-	x 10^{9}/L
49	B-monocytes	< 0.09	0.1 – 0.19	0.2 – 0.29	0.3 – 0.39	> 0.4	-	-	x 10^{9}/L
	Procedure								
50	Urinary catheter	No	Yes	-	-	-	-	-	-

3.2. STUDY I: HOSPITAL-ACQUIRED URINARY TRACT INFECTIONS, MACHINE LEARNING, AND MODEL-SPECIFIC CLINICAL EXPLAINABILITY

This chapter presents the materials and methods of Study I. With the study design, data sources and participants, definitions of HA-UTI outcome, and model design as the backbone through all studies, Study I presents an idea of combining expert knowledge- and machine feature selection and choice of BN models, which enables a discussion of model-specific clinical explainability in the ML model.

3.2.1. FEATURE SELECTION

Expert knowledge draws from recommendations by two infectious disease consultants, supplemented by recent findings on HA-UTI risk factors from the literature [30,41,43,44,105,106]. Given their specialized medical background in microbiology and infectious diseases, these consultants offer essential insights into clinical practices, qualifying them as experts in the field.

An initial feature selection relies on expert-based knowledge, culminating in a *full feature space* that includes admission details, demographics, lifestyle factors, comorbidities, vital parameters, laboratory results, and urinary catheters. The fusion of expert knowledge with tests of marginal independence between the HA-UTI feature and predictive features facilitates the identification of the most significant features within the full feature space. Higher p-values indicate a greater likelihood of feature independence computed within the Hugin framework, the software for constructing the BN models. A feature is included in the *reduced feature space* if expert knowledge deems it significant and the p-value exceeds 0.0005 – a low significance level is chosen due to the large dataset size. Otherwise, the feature only remains part of the comprehensive feature space. Bayesian Network models

A BN model, which concisely represents a joint probability distribution, comprises a qualitative and quantitative part. The qualitative portion is a directed acyclic graph (DAG), with nodes representing random variables and directed edges signifying dependency relationships. On the other hand, the quantitative aspect represents the dependency strengths via conditional probability tables (CPT) for each variable, collectively defining a joint probability distribution. [107]

Eight different BN models are developed and evaluated using four structure learning approaches on full- and reduced datasets. The first method utilizes clinical expert knowledge, referred to as an expert-based clinical model, while the remaining three utilize data learning with varying levels of complexity expressed in terms of allowed conditional dependencies. The expectation-maximization (EM) algorithm [108] learns the model parameters, i.e., conditional probabilities, in all BN models. This involves an expectation step, which includes inference in the underlying model, alternated with a maximization step, where intermediate maximum likelihood parameter estimates are found based on fractional counts derived from the expectation

step [107].

3.2.2. EXPERT-BASED CLINICAL MODEL STRUCTURE

The expert-based clinical BN models are built using a three-step process. First, clinicians identified dependencies between variables, excluding the HA-UTI target node, while adhering to DAG properties. This leads to, e.g., demographics and lifestyle factors being linked to comorbidities, which are further associated with parents to vital parameters and lab results. For example, *BMI* is set as a parent to the *diabetes* node and *alcohol status* to the *liver disease* node, with *diabetes* and *renal disease* as parents to *blood glucose* and *creatinine* levels, respectively.
Second, edges from step one are examined for causality. For instance, an edge is removed if the edge between nodes is due to another condition (like a worsened triage from a disease, which then leads to urinary catheter use, rather than the triage worsening which directly causes the catheter use). However, the edge is kept if it correctly depicts the domain knowledge (like GCS influencing the triage).
Third, a naïve conditional independence assumption is made between predictors and the target HA-UTI node for the reduced feature space, and nodes identified as risk factors by experts had the edge over HA-UTI in the full feature space model (Table 4Structure learning from data
Three approaches for learning BN models from data are compared, whereto the underlying algorithms allow increasing levels of complexity in the model structures. The three approaches are naïve BN [53,107,109], tree-augmented-naïve (TAN) [110], and the PC algorithm [111].

The naïve BN model assumes conditional independence between all features, given the target HA-UTI, resulting in a structure where all edges are directed from the HA-UTI target node to the remaining nodes in the model [53,107,109].
The TAN BN extends the naïve BN by introducing a tree structure over the features (excluding the HA-UTI target variable), in which each node (apart from the root and target node) has two parents [109]. The first step of the TAN algorithm involves developing a weighted, fully connected, undirected graph over the features (excluding the target, HA-UTI), where the weights of the edges correspond to the mutual information between the associated variables, conditioned on the target variable HA-UTI. Subsequently, a maximum weighted spanning tree is established, preserving the edges with the highest conditional mutual information between variables. Since age holds the highest mutual information with HA-UTI (Table 3), it is set as the root node, and all edges in the spanning tree are directed away from age, thereby producing a directed tree structure. Finally, HA-UTI is integrated into the model as a parent to all other nodes [110].
The PC algorithm [107] is a constraint-based method that relies on local conditional independence tests between variables [111]. It starts by forming a fully undirected network. In an iterative second step, it conducts conditional independence tests for all pairs of features, removing any edges between conditionally independent nodes. During this step, it also iteratively searches for V-structures (i.e., a child with two

parents) to set as many directions as possible. Finally, in the fourth step, the remaining undirected edges are oriented to maintain the properties of a DAG and avoid introducing new V-structures [107,111,112].
[11]

3.2.3. EVALUATING THE MODELS

To evaluate the performance of the BN models, a Receiver Operating Characteristic (ROC) curve, an associated AUC, and a confusion matrix are presented. Model-specific explainability of the BN models is discussed in the context of simplicity, transparency, and the ability to capture clinical conditional dependencies.

3.3. STUDY II: HOSPITAL-ACQUIRED URINARY TRACT INFECTIONS, MACHINE LEARNING AND MODEL-AGNOSTIC CLINICAL EXPLAINABILITY

This chapter presents the materials and methods of Study II. With the study design, data sources and participants, definitions of HA-UTI outcome, and model design as the backbone through all studies, Study II distinguishes by the idea behind the feature selection, the training of different ML model types, and the application of SHAP summery- and force plots, which enables a discussion of model-agnostic clinical explainability in the ML model.

3.3.1. FEATURE SELECTION

Two feature selection processes are performed to curtail the impact of potentially extraneous features and optimize ML models' performance, efficiency, and robustness [113], involving both automated χ^2-based selection [114] and manual selection guided by expert knowledge. The automated feature selection is facilitated using the ML environment, scikit-learn (sklearn), version 1.0.2 (https://scikit-learn.org/) [115]. Feature selection based on expert knowledge relies on the literature about risk factors for HA-UTI [30,41,43,44,105,106] and recommendations by clinical experts. The dataset incorporates at least one feature from categories: health socio-demographics,

[11] In step 1 of the PC, a fully undirected graph is constructed. In step 2, a conditional independence test is iteratively performed for all feature pairs, resulting in an undirected network skeleton. In step 3, which happened within the iterating loop in step two, an iterative search for V-structures (e.g., a child with two parents) is used to capture as many directions as possible. In step 4, if some edges remained undirected, they are directed randomly without violating the DAG properties**.** HA-UTI (hospital-acquired urinary tract infection) and GCS (Glasgow Coma Scale)

comorbidities, vital parameters, laboratory results, and urinary catheter usage. Following feature selection, each of the two pared-down datasets contains ten features.

Finally, the dataset is imputed with the most commonly assessed discrete level for the methods that necessitate complete data (NN and LR). The most frequent assessment for each metric pertains to the normal/uncomplicated level, e.g., non-smoking, non-alcoholic, a pulse rate between 50 and 89 beats per minute, an oxygen saturation above 96%, and P-CRP below 7.99 Mg/L, etc. Missing data are encountered in smoking-, alcohol-, and exercise statuses, along with all vital parameters and laboratory test results.

3.3.2. MACHINE LEARNING MODELS

In all, 138,560 unique hospital admissions make up the dataset, which includes 1,877 instances of HA-UTI. 80% of the dataset is used for training the ML models, containing 110,870 (1,484 HA-UTI instances). 20 % of the dataset is used for testing the ML models, containing 27,690 (393 HA-UTI instances) unique hospital admissions. Data are randomly split between the training- and test sets.

Seven different ML models are trained, validated, tested, and compared across the full dataset and our two reduced datasets (Table 5), culminating in the development of 21 ML models. Models such as LR, naïve BN, and DT are employed for their model-specific explainable nature [83]. In addition, a black-box NN and three ensemble methods: RF, AdaBoost (AD), and Gradient Boosting (GB), are also examined for their potential utility in early risk stratification of HA-UTI. The training of all ML models is facilitated by Python's sklearn machine learning package, version 1.0.2 (https://scikit-learn.org/) [115].

The LR models are formulated with L2 regularization, employing the SAGA solver and a maximum of 600 iterations. The DT models utilize the Categorical and Regression Tree (CART) algorithm [116], utilizing the Gini impurity criterion for split quality assessment, optimal split strategy, and an upper limit of eight for tree depth. The NN models are built on a feed-forward architecture, trained via the Adam solver, and limited to 600 iterations. Each NN is comprised of ten hidden units across three hidden layers. The RF models incorporate a maximum of 100 trees (averaging used for decision refinement), the Gini impurity criterion for training, and a maximum tree depth of eight. The AD models employ the AdaBoost-SAMME.R algorithm for boosting, with a limit of 100 trees. The GB models, using the Friedman mean-squared-error for decision-making, also encompass a maximum of 100 trees.

3.3.3. MODEL VALIDATION

A 10-fold cross-validation is performed over the training set for ML model validation and parameter tuning. These parameters, guiding the learning process, are adjusted according to the mean AUC calculated across the ten folds.

3.3.4. TEST OF PERFORMANCE

The ROC curve illustrates the trade-off between sensitivity and 1-specificity for the ML models. An AUC score is also computed to facilitate a comparative evaluation of the model's performance on the test set, which remains concealed during the training phase. To supplement the robustness evaluation of the ML models, the mean AUC and standard deviation (std) are computed from a 10-fold cross-validation on the training set. These results are depicted in parentheses on the ROC curve plots.

3.3.5. SHAPLEY ADDITIVE EXPLANATION

The contribution of a feature within SHAP is determined by its corresponding SHAP value, a measure utilized to quantify a feature's impact on a specific ML model prediction [89]. A summary plot (beeswarm) is used for global model-agnostic clinical explainability, whereas a forceplot is employed for local model-agnostic clinical explainability. These plots are created using Python's SHAP package, version 0.41.0.

The calculation of SHAP values for the models is performed using the KernelExplainer, a form of weighted linear regression [89]. Although its application may be slower due to the circumvention of model-specific assumptions, KernelExplainer serves as a universal explainer, suitable for any ML model type, inclusive of naïve BN, which may be incompatible with other explainers [91].

The summary plot incorporates all 393 admissions wherein patients developed HA-UTI from the test set. A random selection of one patient with HA-UTI and two without HA-UTI is undertaken for the force plot. At the global level, a SHAP summary plot delineates the extent of each feature's positive and negative influence across all predictions, color-coding high and low values, and ranking feature importance based on their contribution. On the local level, a SHAP forceplot elucidates the degree of positive and negative impact each feature has on a specific risk prediction.

3.4. STUDY III: HOSPITAL-ACQUIRED BACTEREMIA AND FEASIBILITY OF MACHINE LEARNING-GUIDED STRATEGY FOR REPLACEMENT OF PERIPHERAL VENOUS CATHETERS

This chapter presents the materials and methods of Study III. With the study design, data sources and participants, definitions of outcomes, and model design as the backbone through all studies, Study III adds a target feature of HAB, another data preprocessing, training of different types of ML models, and addressing the multifaceted challenges of feasibility from a paradigm shift in the PVC replacement strategy when guided by ML models presented in this study.

3.4.1. DATA PREPROCESSING

The study cohort is randomly split into 80% for training and 20% for testing. Due to the scarcity of HAB cases, oversampling is conducted on the training dataset using the RandomOversampler from the imblearn environment [117] version 0.10.1. The OneHotEncoder from Python's sklearn ML package version 1.0.2 [115] encodes the categorical features in both datasets.

The most frequent discrete levels are imputed for categorical features, while the mean value is imputed for all continuous variables. Missing values occur in smoking, alcohol, and exercise statuses, as well as in all vital parameters and laboratory results. To evaluate the robustness and determine a decision threshold for each ML model, a 5-fold cross-validation is performed on the training data.

3.4.2. MACHINE LEARNING MODELS

The design of the ML model enables a hybrid solution focused on distinguishing patients who might significantly benefit from routine replacement of PVC. In contrast, those at a relatively lower risk could proceed with the clinically indicated replacement of PVC. The patients that benefit from routine replacement of PVC may be the same as being at risk of HAB.

A NN, LR, BN, DT, and RF are trained for the purpose of this study using sklearn [115]. A grid search is performed for all ML models to fine-tune the parameters. The NN model, a feed-forward architecture, utilizes the Adam solver, with a hidden layer size of 100, an alpha of 4, and a maximum iteration count set to 700. The LR model, regularized by l2, employs the saga solver with a maximum iteration count of 700. The BN model is trained with the ComplimentNB, which is suitable for a skewed dataset [113]. Based on the CART approach [111], the DT model uses the Gini impurity criterion, the best splitter, a maximum depth of 5, and a maximum of 5 leaf nodes. The RF model, with a maximum of 100 trees that employ averaging to enhance

decision-making, utilizes the Gini impurity criterion and forgoes bootstrapping. Each tree has a maximum depth of 10, with a minimum of 5 samples required to be in a leaf.

In order to enhance the applicability of the ML model by facilitating ML-guided routine replacement of PVC for the most vulnerable patients, several decision thresholds for pinpointing patient admissions at the highest risk of HAB are assessed. Nevertheless, the cost-effectiveness analysis utilizes a decision threshold associated with identifying the top 20% of patients at the highest risk as base case values.

3.4.3. DESCRIPTIVE STATISTICS

The distribution of hospital admissions with and without HAB for train- and test sets are presented. Categorical features are represented in percentages, while continuous variables are expressed as a median with corresponding quartiles.

3.4.4. TEST OF PERFORMANCE

A ROC curve is used to depict the relationship between sensitivity and 1-specificity in the ML models, whereto the AUC enables performance comparison between the models. The mean AUC and standard deviation over a 5-fold cross-validation are included in brackets.
A confusion matrix disclosing a decision threshold, AUC, TP, FP, TN, and FN for decision thresholds identifying 5%, 10%, 15%, 20%, 25%, and 30% of the HAB cases is also reported. The NNH, previously outlined by Laupacis et al. [118], is estimated for each decision threshold and adapted to the context of ML-guided replacement of PVC in preventing HAB. The NNH estimate is calculated by using 70 % of the patients that may are expected to have a PVC [119] for the total positive predicted cases (TP+FP) divided by the number of expected prevented HAB cases, assuming that 14 % of all HAB are related to PVC [45,46], and 86 % of those may be prevented by routine replacement of PVC [120]. Additionally, we report on the cost and effect, which are affected by the performance of the ML models, and include an ICER for each decision threshold.

3.4.5. COST-EFFECTIVENESS ANALYSIS

The construction of a decision tree, distinct from the DT from ML, takes place over various clinical scenarios of HAB occurrence, using TreeAge Pro Healthcare 2021

(https://www.treeage.com/). This process allows a comparison between ML-guided preventive strategy and usual care. The scenarios of the decision tree relying on the ML model-guided preventive strategy align with a test-based approach, previously described by Tamlyn et al. [121], whereas the counterpart encompassing usual care serves as a baseline comparison of the probability of developing HAB.

The decision tree's population relies partially on the performance test results of the ML model with the best performance and partly on estimates accessible from the Business Intelligence and Analysis Unit in the North Denmark Region, such as the attributable cost of a HAB, and hospital lookups like the cost of a PVC and clinicians' salaries. Appendix A presents the methodological considerations in constructing the decision tree. Table 6 presents the early cost-effectiveness analysis's probabilities, costs, and effectiveness estimates.

A deterministic one-way sensitivity analysis, represented through a Tornado diagram, scrutinizes the impact of variations in input parameter point estimates on the ICER [122]. The Tornado diagram orders the inputs based on their effect on the ICER outcome. This order depends on the magnitude of the influence on the output measure when an input parameter changes, with inputs ranked according to their effect size. A red or blue bar signifies the value of one input parameter and how its increase or decrease modifies the ICER result. Costs appear in 2022 values in Danish crowns (DKK), with future costs discounted annually at 3.5 percent [123].

CHAPTER 4. RESULT

This chapter presents the results of the studies conducted in this dissertation. First, the findings of Study I are presented, followed by the findings of Study II and Study III.

4.1. STUDY I: CLINICAL EXPLAINABLE BAYESIAN NETWORK MODELS

Percentages for distributions and missingness for gender, smoking-, alcohol-, and exercise status, comorbidities, and urinary catheters, in addition to median and IQR for age, BMI, vital parameters, and lab results in training- and test set, respectively, for cases with and without HA-UTI, are presented in Table 7 and Table 8.
The performance from the eight BN models is presented in two ROC curves: one for the full feature space and one for the reduced feature space. A confusion matrix is presented for comparison of the BN models. Moreover, the BN models are presented with nodes and edges, enabling introspection of the BN model

4.1.1. MODEL PERFORMANCE

Figure 15 demonstrates the ROC curves over the reduced dataset with only five features (Table 4). Figure 16 presents the ROC curves for the BN models developed over the full feature space (Table 4).

The reduced expert-based clinical BN model is the best-performing model with an AUC of 0.746. The full naïve BN and the reduced TAN BN models follow closely, each achieving an AUC of 0.735. Notably, the reduced expert-based clinical and the full TAN BN models record the least error rates of 32.29 and 29.70 for the decision threshold, respectively (Table 7). It is also worth noting that a model categorically predicting no patient at risk of HA-UTI would have a minimal error rate of 1.37%, as it would correctly identify all non-HA-UTI cases. Nevertheless, such a model would hold no clinical value.

Of the 402 HA-UTI cases in the test set, the reduced naïve BN model identifies 284 risk patients correctly within 24h of admission, achieving the highest TP rates. However, it also records the second-highest false FP, implying a propensity towards aggressive risk classification of patients, accounting for its high error rate of 43.55. Despite reaching the second-highest TP rate, the full expert-based clinical model avoids this high error rate from an elevated FP rate (Table 7). The reduced expert-based clinical and full TAN BN models resulted in the highest TN rates of 18,294 and 19,012 and the lowest FP rates of 8,694 and 8,007, which aligns with their high AUCs (Table 7). The reduced naïve BN model demonstrates the lowest FP-rate of 77, succeeded closely by the full expert-based clinical BN model, which presents an FP-rate of 93. It may be vital to consider that an FN – not initiating preventive strategies for a patient at risk of HA-UTI – is perceived as more critical than an FP, where an unneeded preventive strategy is initiated.

The reduced PC BN model has a fair performance with an AUC of 0.720 (Table 7). However, it is still second-last in the ranking of BN models. The full PC BN model shows a failing performance with an AUC of 0.537, potentially explained by its inability to capture vital HA-UTI dependencies, like age and admission cause. The PC algorithm may be better suited for disease-specific applications, which also impacts the explainability of the BN models, a subject for discussion in the succeeding chapter.

4.1.2. MODEL-SPECIFIC CLINICAL EXPLAINABILITY

BN models have been suggested to convey model-specific clinical explainability by introspection, which unfolds in an inspection of the evidence, the graph, and the reasoning behind the obtained results [62,126].

The naïve BN models do not capture the correct clinical conditional dependencies, which is a direct consequence of conditional independence given the target HA-UTI variable.
Despite the TAN BN model capturing edges from the age node to the admission cause-node, and from the admission cause node to the triage node, the orientation of these edges is dictated by the structure of the tree, which directs all edges away from the root node (Figure 12). Notably, both the naïve- and the TAN BN models reflect simplicity due to the conditional independence assumption (Figure 17 and Figure 18) and because they are restricted to one or two parents, respectively.

The PC algorithm over the reduced feature space (Figure 17) recognizes the admission cause, GCS, and urinary catheter node as parents to the HA-UTI node. This could be because certain admission causes, such as neurological or renal diseases, can be associated with decreased bladder function (increasing the risk of recurrent urine), which may elevate the risk of HA-UTI. Severe admissions may also correlate with an adverse GCS and an increased likelihood of using a urinary catheter to monitor urine output, potentially heightening the risk of acquiring an HA-UTI. However, over the full feature space, the PC algorithm fails to capture edges to more essential features of age and admission cause, as well as the features of triage and urinary catheter, which only form a part of the Markov blanket of the full PC BN model.
Furthermore, the dependencies within the model do not match the relationships expected in the underlying domain, which subsequently affects both the model's explainability and performance (Table 7). For instance, according to d-separation in the full PC BN model, all comorbidities would be excluded from the risk stratification of acquiring HA-UTI if no evidence is provided about the patient's age. This could incorrectly imply that all comorbidities are irrelevant if the age is unknown. A potential explanation for the PC algorithm's inability to capture correct conditional dependencies over the full feature spaces could be attributed to how the admission cause was modeled in our study. General algorithms, including the PC algorithm, for structure learning in infectious domains, may benefit from having a disease-specific dataset for each admission cause.

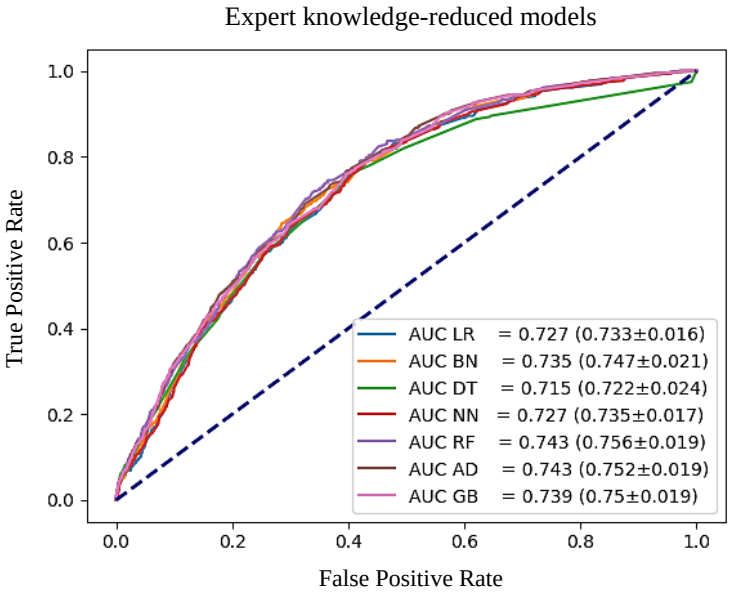
4.2. STUDY II: PROVIDING INSIGHT OF THE REASONING IN A MACHINE LEARNING MODEL

The performance from the 21 different ML models is presented in three ROC curves: one for the expert-knowledge reduced-, one for the χ^2 reduced-, and one for the full feature space. Moreover, A SHAP summary- and force plot is presented to demonstrate global- and local model-agnostic explainability.

4.2.1. MODEL PERFORMANCE

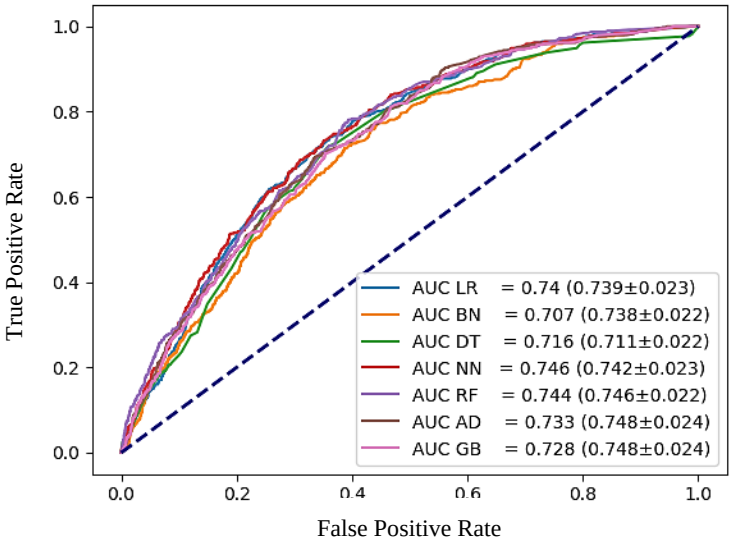
The ROC curves over the 21 ML models (including their respective AUCs from the test and the mean AUC±std resulting from the 10-fold CV) are presented in Figure 19.

Figure 19: The Receiver Operating Characteristic curves for the machine learning models developed over the three datasets[12]

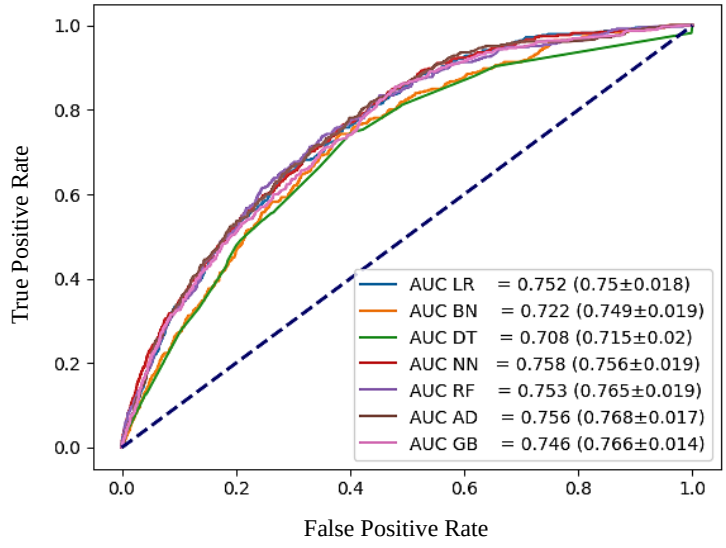


[12] Logistic regression (LR), Bayesian network (BN), and decision tree (DT), Neural Network (NN), Random Forrest (RF), AdaBoost classifier (AD), and Gradient-Boosting classifier (GB), based on the full-, χ^2-, reduced-, and expert knowledge-reduced-dataset.

χ^2-reduced models



Full models



The best-performing ML model on the full dataset (51 features), is the NN, achieving an AUC of 0.758 (0.756,±0.019). Among the models trained on the reduced datasets,

the best-performing is again the NN, based on the $\chi 2$-feature selection method, which achieved an AUC of 0.746 (0.742±0.023).

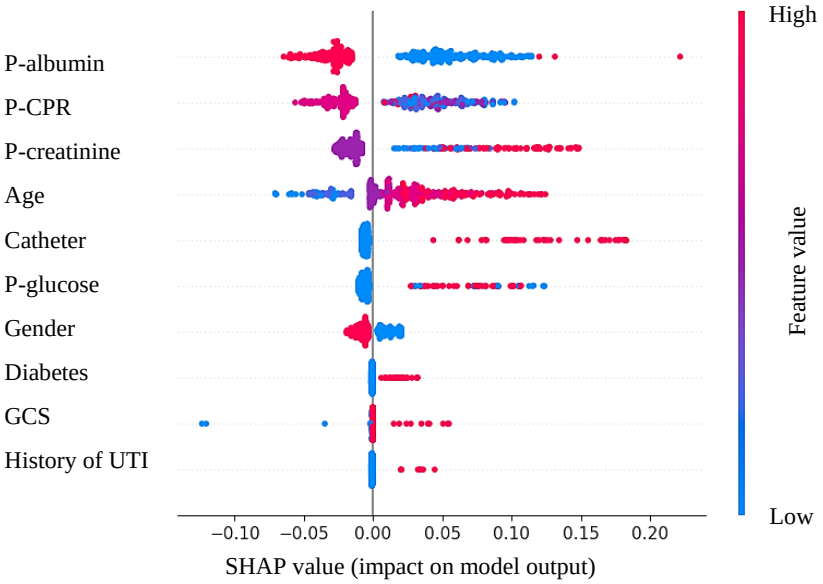
Regarding ML models that may support model-specific explainability, the best performance is from the full dataset and the LR, reaching an AUC of 0.752 (0.750, ±0.018). In contrast, on the reduced datasets, the best-performing model is the LR, again based on the χ^2-feature selection method, which reached an AUC of 0.74 (0.739 ±0.023). Nevertheless, among all ML models that support model-specific explainability, the BN model trained on the dataset reduced by expert knowledge shows the best performance. Notably, the BN trained from the dataset reduced via expert knowledge, which is associated with superior knowledge representation amid the reduced datasets, presents the best performance among all ML models supporting model-specific explainability for this dataset. This BN achieves an AUC of 0.735 (0.747±0.021). Consequently, the BN from the dataset reduced through expert knowledge demonstrates the SHAP methods intended for model-agnostic explainability.

4.2.2. GLOBAL MODEL-AGNOSTIC CLINICAL EXPLAINABILITY

Figure 20 illustrates the SHAP summary plot, using the BN model trained over the expert knowledge-reduced dataset, for the 393 admissions with patients developing HA-UTI.

Figure 20. A summary plot from SHapley Additive exPlanation of the 393 hospital admissions that experienced a hospital-acquired urinary tract infection (using the Bayesian Network model trained from the expert knowledge-reduced dataset) in Study II[13]

[13] The features in the plot are arranged from top to bottom based on their level of contribution to the prediction. A red dot indicates a high feature value, while a blue dot signifies a low feature value. The influence each feature value has on the prediction from a specific hospital admission is plotted on the x-axis, representing each feature's SHAP value.

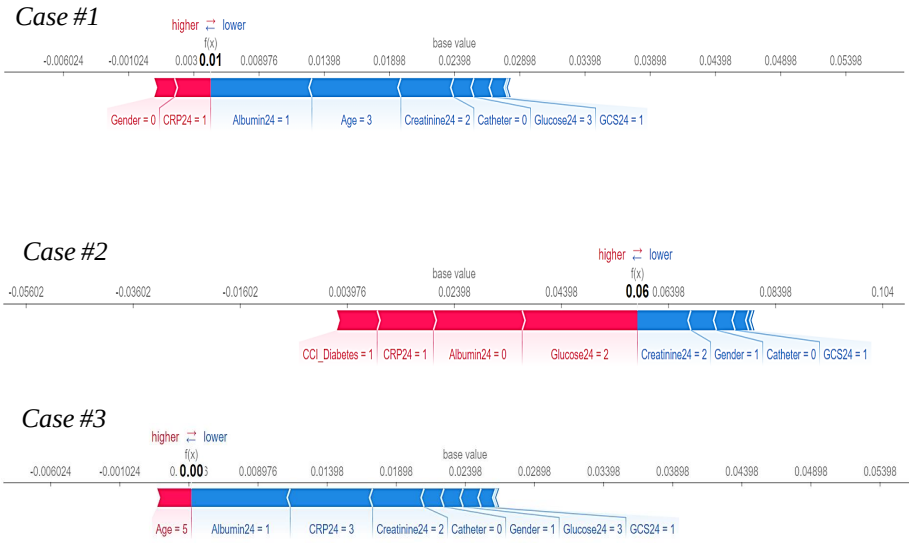


The feature importance is delineated in a top-down manner, which indicates plasma (P)-albumin as the most essential feature in predicting the risk within 24h of admission of HA-UTI. This is followed by P-creatinine, P-CRP, and age. Notably, using a urinary catheter greatly contributes to the likelihood of HA-UTI development. Conversely, diabetes and a history of UTI are found to be the least significant features in estimating HA-UTI risk, despite the slightly heightened probability of HA-UTI in patients with these conditions. Moreover, within the context of the BN model, lower discrete levels of P-albumin are linked with an increased likelihood of HA-UTI. Concurrently, both high and low discrete levels of P-CRP and P-creatinine are associated with a higher risk of HA-UTI. Interestingly, moderately elevated discrete levels of P-CRP and slightly higher levels of P-creatinine seem to mitigate the probability of HA-UTI. Lower age is associated with a lower risk of HA-UTI, while higher age corresponds to an increased risk. The risk of HA-UTI is slightly higher in females compared to males. Finally, high discrete levels of GCS (implying lower actual GCS scores), diabetes, and a history of UTI might slightly elevate the risk of HA-UTI, albeit these features typically provide minimal contributions.

4.2.3. LOCAL MODEL-AGNOSTIC CLINICAL EXPLAINABILITY

The SHAP forceplot for a trio of patients – two without HA-UTI and one with HA-UTI – selected randomly, with their respective data processed through the BN model derived from the expert knowledge dataset, are presented in Figure 21.

Figure 21. Forceplots from SHapley Additive exPlanation of the three randomly chosen patients, two without- and one that experienced a hospital-acquired urinary tract infection (using the Bayesian Network model trained from the expert knowledge-reduced dataset), in Study II[14]



In the SHAP forceplot, the base value represents the model's output when no feature information is available. SHAP values, on the other hand, indicate how each feature influences the predicted outcome (f(x)) based on observed data. Features that increase the predicted value f(x) are depicted in red, whereas those that decrease f(x) are in blue. The bar size in the forceplot corresponds to each feature's SHAP value and is arranged according to its contribution, which relies on the specific discrete level

---

[14] Case #1, a female patient in her 50s. At admission, high P-glucose, normal kidney function, moderate P-CRP elevation, and absence of a urinary catheter contributed to a lowered HA-UTI risk of 1.0% (compared to a 2.40% baseline). The Bayesian Network (BN) model correctly predicted actual absence of HA-UTI, marking a true negative.

Case #2, a male patient in his 40s with diabetes. Moderately high P-glucose and P-CRP levels at admission led the BN model to predict an elevated HA-UTI risk of 6.0%, compared to a 2.4% baseline. Normal P-albumin level, male gender, normal GCS score, proper kidney function, and absence of a urinary catheter helped lower this risk. The BN model accurately predicted clinical HA-UTI presence, marking a true positive.

Case #3, a male patient in his 70s. Despite high age, the BN model predicted a 0.0% HA-UTI risk, lower than the 2.4% baseline. High P-glucose and elevated P-CRP levels at admission, alongside a high P-albumin level, likely initiated an antibiotic treatment, preventing HA-UTI development. The absence of a urinary catheter further reduced the risk. The BN model correctly predicted clinical HA-UTI absence, marking a true negative.

(Table 3). Only the features with SHAP values exceeding a predetermined threshold (set by SHAP) are displayed in the SHAP forceplot.

4.3. STUDY III: FEASIBILITY OF MACHINE LEARNING-GUIDED REPLACEMENT STRATEGY OF PERIPHERAL VENOUS CATHETER

Descriptive statistics present the distributions of cases with- and without HAB between the training- and test sets. The performance of the different ML models is presented in ROC curves, whereto a confusion matrix also describes the NNH and ICER resulting from different decision thresholds, hence assessing different feasibility scenarios.

4.3.1. DESCRIPTIVE STATISTICS

Out of 138,588 hospital admissions that met the inclusion criteria, only 367 (0.26 %) resulted in a HAB. After dividing the data, the training set consisted of 110,871 admissions, including 298 HAB cases, while the test set consisted of 28,017 admissions, including 69 HAB cases,.

In the training set, the 30-day mortality rate for patients who did experience a HAB is 5.32%, while for those who did experience HAB, it is significantly higher at 26.89%. The median length of stay for non-HAB patients is 2.2 days [1.0 - 4.8]. Conversely, for patients with HAB, the median stay extended to 21.2 days [11.0 - 37.7]. The distribution of admission causes, triage, and season for both HAB and non-HAB cases in the training and test sets is presented in Tables 8, -9, and -10, respectively.

4.3.2. MODEL PERFORMANCE

The RF model outperformed the others, achieving an AUC of 0.820. Table 11 provides the confusion matrix for several decision thresholds—5%, 10%, 15%, 20%, 25%, and 30%. This allows for a comparative analysis of these thresholds, considering the AUC, TP, TN, FP, FN, NNH, cost, effect, and ICER across the five ML models.

The study shows that as more patients are tagged as at risk, the NNH rises linearly, reflecting the growth in FP compared to TP. Conversely, the ICER follows a logarithmic trend, with high ICER for low decision thresholds reducing for higher thresholds. This might be due to the high cost of not preventing HAB, indicating that the effect gained outweighs the associated costs when routinely replacing PVC. Therefore, balancing NNH and ICER may optimize the practicality and benefits of ML risk prediction and PVC replacement every 96 hours. A threshold of 20% may strike a reasonable balance in identifying patients who might benefit from routine PVC replacement.

4.3.3. FEASIBILITY AND COST-EFFECTIVENESS

The ML model is more effective at reducing HAB-related deaths per patient, averaging a death probability of 0.000629 compared to 0.000041 for standard care. Nevertheless, it is also more costly per patient, with an average cost of DKK 4,009.49 against DKK 3,950.60 for standard care. The ICER translates to DKK 1,440,495.00 per preventable HAB-related death using the RF model and a 20% risk threshold.

The Tornado diagram shows that ICER is most sensitive to the number of patients screened annually, the anticipated duration of ML model use, and the upfront costs (purchase, implementation, development, maintenance, training) of the ML model. The Tornado diagram is shown in Figure 4.

CHAPTER 5. DISCUSSION

This chapter presents discussions of the findings of this dissertation. First, the findings of Study I are discussed, followed by a discussion of the findings of Study II and Study III. The last section of the chapter describes the limitations and strengths of the three studies.

5.1. STUDY I: ENABLING MODEL-SPECIFIC EXPLANATION

The BN models over the reduced feature space aim to promote simplicity by fewer nodes and edges than models over the full feature space. Of notice, integrating expert-based knowledge in both feature selection and model development may promote the adaption of the applications of ML models in a daily clinical routine [63,64], in addition to enhancing the BN model performance. BN models convey explainability by allowing introspection of the evidence, the graph, and the probabilistic reasoning behind the obtained results [62,127], which may be further associated with promoting transparency, hence model-specific clinical explainability.

5.1.1. INSPECTION OF THE EVIDENCE, THE GRAPH, AND THE REASONING BEHIND THE OBTAINED RESULTS

Study I compared BN structures developed from manual expert knowledge, naïve, TAN- and the PC algorithm. However, other approaches could also have been considered. For instance, Jakobsen et al. [128] trained the structure of a BN from the Greedy-search-and-score algorithm [129] for predicting mortality rates for patients with diabetes in the intensive care unit and compared the performance to existing Acute Physiology And Chronic Health Evaluation II [130]. While the greedy-search-and-score algorithm reached remarkable performances in mortality prediction [128], the TAN algorithm has traditionally reached great performances within the infectious domain [109], favoring this approach for Study I. Also, the perspectives of finding V-structures in the PC algorithm [111] were considered prone to capture correct dependencies with proper inference in the structure, whereto this was also the method of choice for Study I. However, the PC algorithm over the full feature space reached a structure associated with a failing performance (Tale 7) with an unfortunate large CPTs, e.g., the age node with many parents (Figure 18). In this scenario, a greedy-search-and-score algorithm would have allowed for setting a constraint of maximum parents for the nodes, which would solve the issues of large CPTs from the PC algorithm.

Drawing from studies that highlight causality as a critical factor for explainability in ML[59,62,85,131], BN models may be a proper model choice to represent accurate conditional dependencies, thereby enhancing model-specific explainability. However, a small trade-off persisted between achieving explainability and reaching optimal performance for the expert-derived clinical BN. For instance, the nodes for age, triage level, and urinary catheter usage were initially expected to be parents of the HA-UTI node (not the other way around). In this scenario, higher age may be associated with a weakened immune response; adverse triage indicates severe patient condition, and urinary catheter use is linked to a higher risk of HA-UTI. However, reversing the edge between age, the most significant predictive feature for HA-UTI (Table 4), and the HA-UTI node reduced the AUC to 0.743 from 0.746 (Table 5). When all edges between the predictive features and the target HA-UTI node were reversed, aligning with anticipated domain relationships, the AUC in the performance test of the reduced expert-based clinical model dropped significantly to 0.697. These results suggested that our expert-based clinical BN model's performance depended on the naive assumption of independent relationships between predictive features and the target HA-UTI node. This understanding would limit the utility of techniques such as Noisy-or and divorcing [107], which aim to reduce the number of parents but would not change the orientation of the edges. Consequently, we accepted a trade-off between performance and model-specific explainability for our expert-based clinical models to reach higher AUC with the price of not reflecting entirely accurate dependencies to the target node. Of notice, this introspection was possible for BN models, hence supporting model-specific explainability.

5.1. STUDY II: PROVIDING A MODEL-AGNOSTIC EXPLANATION

This study delves into the training, validation, testing, and comparison of various ML models for identifying patients within 24h of admission at risk of HA-UTI. Additionally, the SHAP analysis facilitated clinical explainability in these ML models.
The NN training over the full dataset was the best-performing ML model. ML models were also trained over two reduced datasets, one from manual expert knowledge and one automated data-driven method, reaching fair performances while only relying on the ten most significant features. The SHAP-method further amplified the BN model's global- and local model-agnostic explainability, providing clinical reasoning behind the ML model's risk prediction.

5.1.1. THE IMPORTANCE OF THE CLINICAL FEATURES

The two feature selection methods (manual and automatic) agreed on the following four features as most essential for risk stratifying within 24h of admission for HA-

UTI: age, urinary catheter, P-CRP, and P-albumin (Table 5). While age and urinary catheter are well-established HA-UTI risk factors [41,43], the inclusion of P-CRP and P-albumin may reflect the generic clinical nature of these features and the ML model design. For instance, P-CRP [106] and P-albumin [105,106,132] are broad biomarkers, indicating both chronic and acute diseases, thus suggesting a connection to the patient's general health status at admission, contributing to the HA-UTI risk assessment within 24h of admission (Figure 10).

In the χ^2-based feature selection, high χ^2values for features such as cause of admission, a total of comorbidities, and B-lactate suggest a significant difference between expected and observed counts. This implies that these features have a greater dependence and may be better predictors of HA-UTI.
On the other hand, the expert knowledge-reduced dataset includes gender, diabetes, UTI history, GCS, and P-glucose, which are absent in the χ^2-reduced datasets (Table 5). Each of these features is associated with recognized risk factors for HA-UTI [41,43,133]. High P-creatinine levels, indicative of impaired kidney function, are also recognized as a UTI risk factor [105,134]. The GCS assesses the patient's mental status and may be influenced by chronic or acute conditions [135] associated with an increased risk of UTIs. P-glucose levels can indicate diabetes, stress-induced conditions, insulin sensitivity influenced by fever, or kidney damage, which could elevate the risk of HA-UTI. Proper glycemic control may be crucial for preventing HA-UTI [136]. It is worth noting that while the χ^2-reduced datasets lean towards laboratory results, the expert knowledge dataset incorporates a broader range of health socio-demographic and clinical data (Table 5).

5.1.2. THE BALANCE OF PERFORMANCE AND CLINICAL EXPLAINABILITY

In comparison to study II, Møller et al. [67] reported that their highest-performing ML model for predicting HA-UTI risk within 48 hours of admission was a NN model, achieving an AUC of 0.770. However, because of the opaque nature of NN models, they advocated for the more interpretable DT model, which attained an AUC of 0.709. The best-performing DT model in Study II was derived from the χ^2-reduced dataset and achieved an AUC of 0.716. However, LR- and BN models,- associated with supporting model-specific explainability, outperformed DTs on all datasets (Figure 19). Model-agnostic explainability at both global- and local levels was demonstrated to enhance the clinical explainability of the ML models further.

The approach in Study II aligns with the approach of Deshmukh et al. [137] and Stenwig et al. [91], who utilized SHAP forceplots and SHAP summary plots for local and global model-agnostic interpretability in predicting mortality in intensive care units. Moreover, a recent Danish study by S. Lauritsen et al. [138] employed the same SHAP methods for the early detection of acute illness. Notably, the SHAP library provides many visualization techniques, offering various visual explanations for

global and local model-agnostic approaches. Future research could explore how different plots contribute to the comprehension of clinical reasoning behind ML models' outputs or how SHAP explanations might differ from other model-agnostic methods, such as Local Interpretable Model-Agnostic Explanations (LIME) [61].

Finally, the SHAP method can be instrumental in assessing the potential clinical implications of a machine learning model's recommendations. For instance, in the case of patient 3 (Figure 21), a high P-CRP level results in a blue bar in the SHAP forceplot, indicating a negative influence on the prediction of HA-UTI risk. This may imply that the patient underwent some form of treatment, such as antibiotic therapy, during their hospital stay. Therefore, a high P-CRP level might be construed as a surrogate marker for 'administered antibiotic therapy' rather than indicative of a 'higher risk of infection', which would explain the negative impact manifested in the SHAP forceplot. Of note, the data on antibiotic treatment was not available.

5.1. STUDY III: A HYBRID REPLACEMENT STRATEGY FOR PERIPHERAL VENOUS CATHETER

The ML models in this study were able to predict the risk of HAB within 24 hours of admission. By setting a decision threshold to identify the top 20% of patients at the highest risk of HAB, 72.46% (sensitivity) of all HAB cases were identified within the first 24h of admission using an RF model (AUC of 0.820). Suppose these 20 % of the cohort were identified to benefit from routine replacement of PVC to avoid HAB, then the NNH was 766 for avoiding one HAB. Of notice, this was significantly better than recent research suggesting an NNH of approximately 3500 for a strategy entirely dependent on routine replacement of PVC for HAB prevention [120].
Considering the costs, the difference between standard care (DKK 3,950.60) and the ML-guided routine replacement of PVC every 96 hours [9] (DKK 4,009.49) is a minor increase of DKK 60.13 per admission. The cost-effectiveness analysis demonstrated an ICER of DKK 1,440,495.00 per avoidable HAB-related death for the RF model, given the decision threshold that labels the top 20 % as HAB-risk patients. Ultimately, the cost-effectiveness of this ML model hinges on decision-makers willingness to pay for each avoided HAB-related death.

Patients who were male or of an older age were found to have an increased risk of developing hospital-acquired bacteremia (HAB). Those admitted due to neoplasms or diseases of the digestive system also had a higher percentage of HAB occurrences. This could potentially reflect the higher percentages of patients with comorbidities, e.g., cancer, leukemia, and lymphoma, amongst those who developed HAB. Triage indicated that patients who developed HAB had a higher severity of illness than those without HAB. Seasonality also appeared to play a role, with slightly increased risks observed for patients admitted during the fall or winter compared to those admitted in spring or summer. Regarding clinical measures, P-creatinine and P-CRP levels

showed a significantly higher median (and quartiles) for HAB cases, indicating that these measures may be key risk indicators.

5.1.1. RISK STRATIFICATION OF HOSPITAL-ACQUIRED BACTEREMIA

The ML models in this study were designed to support decision-making within the first 24 hours of hospital admission, enabling the early implementation of preventive strategies against HAB. This approach aligns with the work of Parreco et al. [139], who used LR, GB, and NN to predict central line-associated bloodstream infection based on socio-demographic and clinical features from the MIMIC-III database [140]. While their LR model had the highest AUC of 0.722, it lacked sensitivity, a limitation they acknowledged. Furthermore, they did not discuss how the model's low sensitivity might affect clinical intervention or cost-effectiveness.

Meanwhile, Mahmoud et al. [69] used several ML models, including NN, LR, BN, DT, RF, and SVM, for the early identification of bacteremia. Their LR model, using Synthetic Minority Over-sampling Technique for imbalanced data, achieved the highest sensitivity at 31%, with a specificity of 73%. The RF model in Study III, in contrast, achieved a sensitivity of 72.46% and a specificity of 76.34% when using a decision threshold to identify the top 20% of patients at high risk for HAB and applying random oversampling.

Garnico et al. [72] achieved an impressive AUC of 0.93 with a sensitivity of 87.4% using SVM, RF, and k-nearest neighbor in their work predicting HAB. However, their dataset was relatively small and consisted of 117 features, potentially making it susceptible to overfitting. Despite recognizing the potential cost benefits of ML for predicting HAB, they did not consider cost-effectiveness in their model.

Finally, Beeler et al. [141] trained an LR and an RF for real-time prediction of CLABSI, providing a new daily prediction and achieving an AUC of 0.79 for LR and 0.82 for RF. While their work specifically targeted CLABSI rather than HAB, the concept of providing updated risk predictions every 24 hours could potentially be incorporated into our models. However, within 24 hours of admission, predisposing factors for HAB, such as male gender, higher age, or admission due to neoplasms or diseases of the digestive system, would already be captured, with only clinical features subject to real-time updates.

5.1.2. CLINICAL INDICATED- AGAINST ROUTINE REPLACEMENT STRATEGIES

Whether to replace PVC based on clinical indication or a routine basis is widely debated in the clinical context [119,125,142–146]. Some traditional, often older, studies suggest that replacement based on clinical indication is superior, as it does not significantly alter the incidence of HAB while reducing economic cost, e.g., being less time-consuming [119,142,143]. However, more recent studies suggest a shift in

the best-practices, proposing that routine replacement may prevent more cases of HAB [125,146]. Nonetheless, the feasibility of routine replacement may be questioned due to the associated increase in time and cost and the challenge of implementing this in clinical practice, especially considering the low incidence of HAB [125,146]. A study by Meirson et al. [120] estimated an NNH of around 3500 for a strategy that relies entirely on routine PVC replacement to prevent HAB. Interestingly, Study III proposes a hybrid solution that can significantly reduce the NNH by targeting only those patients at high risk of HAB who might benefit from routine PVC replacement, a strategy enabled by ML risk prediction.

The estimated cost of an admission involving HAB in the North Denmark Region was DKK 57,542 (€7,692.78), compared to the usual cost of an admission without HAB at DKK 3,782 (€505.61), resulting in an attributable HAB-related cost of approximately DKK 53,760 (€7,187.17). This was notably lower than the mean costs reported by Riu et al. [36] in a 2016 study from Spain, which were DKK 193,664.68 (€25,891) for admissions with HAB and DKK 50,490 (€6,750) for usual care, leading to a higher attributable cost of approximately DKK 143,174.68 (€19,141). Other studies have estimated even higher HAB-related costs, ranging from DKK 197,789 ($29,000) [35,147] to DKK 327,360 ($48,000) [148]. The lower estimated costs in this study could lead to underestimating the potential savings from initiating routine replacement of PVC guided by ML risk prediction. Moreover, it may lead to overestimating the ICER of avoiding a HAB-related death due to the low attributable cost of HAB.
The three main factors influencing the overall cost were the number of patients screened each year, the time frame from which the model is expected to be used, and the estimated costs of implementing the ML model (Figure 24). The annual cost of implementing, maintaining, and training staff for the ML model was estimated to be DKK 750,000 (€100,267.38), an estimate with considerable uncertainty. Higher estimated costs would result in higher attributable costs and hence a higher ICER, whereas lower costs would result in the opposite. Notably, factors such as increased time consumption, higher salaries, and using more PVCs did not significantly impact the ICER. Also interestingly, the number of FN cases had a greater influence on the ICER than the number of FP cases, indicating that the sensitivity of the ML model has a more significant effect on the cost-effectiveness than the specificity (Figure 24).

5.2. STUDY LIMITATIONS AND STRENGTHS

In Study I, only one type of ML model was explored for risk stratification of HA-UTI, yet from different structure learning methods, and for the additional purpose of elucidating model-specific clinical explainability. Study II presented a horizontal view of evaluating different types of ML models in risk stratification of HA-UTI and described both global- and local model-agnostic clinical explainability, but did not delve into model-specific explainability like Study I. Study III changed its purpose to risk stratification of HAB, also from different ML models, but without considering

expert knowledge or explainability, yet to provide insight on cost-effectiveness and feasibility. While each study was limited to its respective purpose, they may complement each other by shedding light on clinical explainability for risk stratification of HAI, which also includes an assessment of the feasibility for such a purpose. Also, they were subject to a similar model design, which intends to promote simplicity. In this scenario, predicting adversities that happen >48h after admission within the first 24h can be challenging as various factors might influence the outcome between prediction and the occurrence of the adverse event (Figure 10). However, promptly identifying patients who are at risk is of substantial clinical importance, enabling immediate preventive actions, such as reducing the length of potential catheterization or assisting in the early suspicion of HA-UTI, or deciding on replacement strategies for PVC in avoiding HAB. Notably, the early guidance by risk assessments in the studies presented in this dissertation may be close in spirit to the purpose of MiBAlert in Denmark, suggesting that a patient who has experienced a multi-resistant pathogen within a particular time window (varying between pathogens) could be subject to isolation [149]. Providing these early risk assessments may enable timely and targeted preventive measures for HAI in the future.

While the HA-UTI- and HAB case definitions utilized in this research are deemed highly reliable [26], the included comorbidities rely solely on ICD-8 and ICD-10 codes (Table 2), sometimes used for reasons other than recording disease history. Consequently, the dataset may occasionally encounter a negative case, such as diabetes or a history of UTI, that should have been identified as positive. This oversight could have improved performance or explained a more accurate underlying pattern. Additionally, the dataset did not incorporate information regarding antibiotic therapy, an established risk factor for HA-UTI [41] and HAB [34]. Other therapies, like PVC for Study III, could have also influenced the outcomes. For instance, whether patients had PVC could have reduced the dataset by approximately 30% [119]. However, it might have also eliminated patients that could introduce noise to the data. If these data become available in the future, the performance of the ML models presented in this dissertation may likely improve.

For the NN and LR models in Study II, which require a complete dataset, missing data were handled by imputing the most frequent value for each feature. For Study III, mean values were imputed for the continuous variables. However, in a scenario where clinical justification is required, this approach could result in faulty conclusions for HA-UTI risk prediction, as it relies more on approximations than precise clinical measurements. This shortcoming could, e.g., affect the results of the SHAP analysis of Study II. Conversely, the BN and tree-based ML models are not dependent on datasets with imputed values. Clinicians should judiciously interpret future applications of clinical explainable ML models in infection control, as they may suggest decisions from another underlying and clinically relevant pattern than those initially intended.

In Study III, the feasibility of ML models was assessed in an early cost-effectiveness analysis. However, feasibility can be a multifaceted challenge that extends beyond

purely economic considerations and NNH. For instance, the time required for routine replacement of PVC might correlate to allocating staff hours from a constrained time resource in a high-pressure clinical environment. This allocation could even differ between hospitals. In such circumstances, changing preventive strategies might not be feasible due to insufficient staff hours, even if the overall cost is not a problem. A potential solution to this challenge might involve lowering the decision threshold of the ML models highlighted in this study, such as 5-10% (Table 11). This would mean fewer patients would be designated at-risk, reducing the associated NNH. As a result, the number of TP might decrease due to a less time-consuming strategy, but with a higher number of FN and an increased ICER for preventing HAB-related deaths. This flexibility in setting different decision thresholds in a clinical context could potentially enhance the feasibility of an ML-guided strategy for preventing HAB.

Lastly, while the ML models presented in Study I, -II, and -III may potentially enhance timely and targeted preventive measures of HAI in the future, they might only be considered in balance with the remaining eight core components of IPC guidelines suggested by WHO [19–22]. This perception is core to understanding the ML models fit into the clinical routine.

CHAPTER 6. CONCLUSIONS

This dissertation evaluates the use of clinical explainable ML models for risk stratification within 24h of admission for HA-UTI and the feasibility of an ML-guided replacement strategy of PVC in avoiding HAB. This enables timely and targeted preventive and therapeutic strategies for patients in most need.

Study I and -II presented seven different types of ML models in risk stratification of HA-UTI. Predictive features include data on admission detail, demographics, lifestyle factors, comorbidities, vital parameters, laboratory results, and procedures. Probabilistic BN models were used to delve further into model-specific clinical explainability, and the SHAP method was used to elucidate model-agnostic clinical explainability. Three types of feature selections were conducted, one with concatenating expert-based knowledge and p-values on the test of marginal independence, one with automated χ^2, and one relying solely on expert knowledge. All feature selection approaches recognized age and urinary catheter as essential features for risk stratification of HA-UTI within 24h of admission. Two feature selection approaches identified admission cause, GCS, P-albumin, and P-CRP as the most essential features. One feature selection approach recognized gender, triage level, diabetes, History of UTI, B-thrombocytes, P-glucose, B-hemoglobin, P-creatinine, P-lactate, and P-monocytes as most important for risk stratification of HA-UTI within 24h of admission.
Seven out of eight BN models in Study I reached a fair AUC between 0.720 and 0.746 while supporting model-specific. The reduced expert-based clinical BN model reached the highest AUC in Study I and was further associated with the best reflection of dependencies in the domain. The different ML models in Study II reached an AUC between 0.707 and 0.758. The NN over the full feature space reached the highest AUC. However, to promote simplicity with fewer features from a reduced dataset associated with better knowledge representation and from an ML model that enables model-specific explainability, the BN model developed over the expert-knowledge reduced feature space was suggested for further SHAP analysis (AUC of 0.735). The SHAP summary plot visualized a ranked contribution of each feature, dependent on the feature values, for the part of the cohort that experienced an HA-UTI. The SHAP forceplot visualized the impact of the features of the risk stratification of a single patient.
Study I and -II contributes to the ongoing debate on the balance between performance and explainability between ML models [62,78,84,86,89,137,150,151] – at least for model-specific explainability for risk stratification within 24h of admission for acquiring HA-UTI. Interestingly, no significant compromise in performance was found between ML models that support model-specific explainability and those that do not. This aligns with the recommendation of Rudin [83] of using ML models that facilitate model-specific explainability, particularly in sensitive healthcare applications.

Study III, which shares the model design with Study I and -II, introduced five ML models that identify patient admissions with the highest risk of HAB within the first 24h of admission. These patients are most likely to benefit from the routine replacement of PVC. The remaining cohort may continue with usual care based on clinical indications, which could be a slightly more time-saving approach [142]. The ML models reached AUCs between 0.69 and 0.82 in the test, with the best-performing ML model being the RF. From a decision threshold that labels the top 20 % as HAB-risk patients, the NNH of relying on a routine replacement strategy for these risk patients of PVC was 766 (compared to the literature suggested 3500 over the entire cohort), with an ICER of DKK 1,440,495.00 for avoiding a HAB-related death. Notably, the concepts put forward in Study III could suggest a potentially significant shift in the traditional debate of routine versus clinically indicated replacement of PVC. Instead, we might move towards a risk-dependent clinical strategy to prevent a higher number of HABs, guided by the data-driven mechanisms of ML models. This method may also be feasible when considering the associated NNH and costs.

CHAPTER 7. FUTURE PERSPECTIVES

The future direction for this dissertation's clinical explainable ML models could be to reach implementation. Inspiration for this could be taken from the realm of responsible ML in healthcare [64,92]. However, further research is required to fully grasp the role of ML, its clinical explainability, and its feasibility in future infection control for HA-UTI and HAB. Ultimately, the aim is to benefit future patients by mitigating these adverse outcomes.

Perspectives of not restricting the model design to a simple risk stratification approach within 24h of admission may invite ideas of dynamic BN models to capture a potential development in some features over time. It would also be interesting to see how this may impact the model-specific clinical explainability when the model complexity may increase from multiple time slices in a dynamic setting. Also, other model-agnostic approaches, e.g., LIME or counterfactual explanations, may be explored and compared to SHAP at a research level, but also how these explanations may – or may not – assist the clinical in interpreting the suggestions for their given daily routine.

The costs associated with HAB are based on Danish data from the Business Intelligence and Analysis Unit in the North Denmark Region and hospital lookups, and they might be underestimated. This was, e.g., due to considering only within-hospital costs and not accounting for readmission costs, potential rehabilitation, etc. Likewise, the effect was based on avoiding HAB death, but other effects could also have been encountered, such as HAB-related ventilator dependence and renal failure [152] or quality-adjusted life years (QALY). Future studies on the cost-effectiveness of preventing HAB could benefit from more comprehensive definitions of what should be included in assessing HAB costs. For instance, it might consider types of antibiotic therapy, more medical staff involvement leading to increased salary costs, potential rehabilitation, etc. This broader approach might also help elucidate why the costs in this study could be underestimated.

Lastly, the processes of developing and implementing clinically explainable ML models for HAI could draw inspiration from other fields, e.g., the interplay between innovation, strategic idea portfolios, and XAI. In this scenario, inspiration may be found in the work of Jakobsen et al. [153]. For instance, while strategic idea portfolios relate to ensuring the innovation process of constructing the best ideas, ensuring absorptive capacity and pursuing blindspotting, these perspectives may be similar to the hopes of ensuring the process by the steps of responsible ML. Also, similarities between strategic planning- and explainability methods, such as the Forcefield [154] and SHAP summary plots [89], may inspire how others use the weights of feature contributions to reach informed decision-making. Future studies on clinical explainable ML for HAI may benefit from pursuing multidisciplinary approaches across different fields of expertise to explore how ML models may fit the remaining

eight core components of the IPC guidelines proposed by the WHO [19–22], hence increasing our understanding of how ML models may fit the daily clinical routine.

LITERATURE LIST

[1] Garner JS, Jarvis WR, Emeri TG, et al. CDC definitions for nosocomial infections. Am J Infect Control. 1988;

[2] World Health Organization W. Prevention of hospital-acquired infections. 2002;

[3] Hensley BJ, Monson JR. Hospital acquired infection. Cambridge Handb Psychol Heal Med Second Ed. 2014;33:736–738.

[4] World Health Organization. WHO Health care-associated infections FACT SHEET. 2010;

[5] Zingg W, Holmes A, Dettenkofer M, et al. Hospital organisation, management, and structure for prevention of health-care-associated infection: A systematic review and expert consensus. Lancet Infect Dis. 2015;15:212–224.

[6] Yallew WW, Kumie A, Yehuala FM. Risk factors for hospital-acquired infections in teaching hospitals of Amhara regional state, Ethiopia: A matched-case control study. PLoS One. 2017;12:1–11.

[7] Raoofi S, Kan FP, Rafiei S, et al. Global prevalence of nosocomial infection: A systematic review and meta-analysis. PLoS One. 2023;18:1–17.

[8] Gubbels S. The development of a national surveillance system for hospital-acquired infections in Denmark The Hospital Acquired Infections Database - HAIBA. 2016.

[9] Glied S, Cohen B, Liu J, et al. Trends in mortality, length of stay, and hospital charges associated with health care–associated infections, 2006-2012. Am J Infect Control. 2016;44:983–989.

[10] Condell O, Gubbels S, Nielsen J, et al. Automated surveillance system for hospital-acquired urinary tract infections in Denmark. J Hosp Infect. 2016;93:290–296.

[11] Statistik D. Befolkningens udvikling. 2021.

[12] Statsrevisorerne (Rigsrevisionen). Rigsrevisionens beretning om forebyggelse af hospitalsinfektioner. 2017.

[13] Pedersen KM, Hansen MB, Hans Jørn Kolmos. Hospitalsinfektioner har store menneskelige og økonomiske konsekvenser. Ugeskr Laeger. 2021;

[14] Sundhedsdatastyrelsen. Udvalgte nøgletal for sygehusvæsenet og praksisområdet 2009-2018. 2020;

[15] Cardoso T, Ribeiro O, Aragão I, et al. The Impact of Healthcare-Associated Infection on Mortality: Failure in Clinical Recognition Is Related with Inadequate Antibiotic Therapy. PLoS One. 2013;8:1–8.

[16] Naimi TS, Ledell KH, Como-sabetti K, et al. Comparison of Community- and Health Care – Associated Staphylococcus aureus Infection. 2023;290.

[17] Friedman ND, Kaye KS, Stout JE, et al. Health care-associated bloodstream infections in adults: A reason to change the accepted definition of community-acquired infections. Ann Intern Med. 2002;137:791–797.

[18] Harbarth S, Sax H, Gastmeier P. The preventable proportion of nosocomial infections: An overview of published reports. J Hosp Infect. 2003;54:258–266.

[19] World Health Organization. Guidelines on core components of infection prevention and control programmes at the national and acute health care facility level. 2016.

[20] Storr J, Twyman A, Zingg W, et al. Core components for effective infection prevention and control programmes: New WHO evidence-based recommendations. Antimicrob Resist Infect Control. 2017;6.

[21] World Health Organization. MINIMUM REQUIREMENTS for infection prevention and control programmes. 2019.

[22] Puro V, Coppola N, Frasca A, et al. Pillars for prevention and control of healthcare-associated infections: an Italian expert opinion statement. Antimicrob Resist Infect Control. 2022;11:1–13.

[23] Saklani P, Krahn A KG. Urinary Tract Infection. 2013;

[24] World Health Organisation. WHO guidelines on Hand Hygiene in Health Care. World Heal Organ. 2017.

[25] Gould IM. Antibiotic policies to control hospital-acquired infection. J Antimicrob Chemother. 2008;61:763–765.

[26] Michalopoulos A, Falagas ME, Karatza DC, et al. Epidemiologic, clinical characteristics, and risk factors for adverse outcome in multiresistant gram-negative primary bacteremia of critically ill patients. Am J Infect Control [Internet]. 2011;39:396–400.

[27] Paul M, Kariv G, Goldberg E, et al. Importance of appropriate empirical antibiotic therapy for methicillin-resistant Staphylococcus aureus bacteraemia. J Antimicrob Chemother. 2010;65:2658–2665.

[28] Sips ME, Bonten MJM, Van Mourik MSM. Automated surveillance of healthcare-associated infections: State of the art. Curr Opin Infect Dis. 2017;30:425–431.

[29] Chen LF, Ou TY, Teng SO, et al. Hospital-acquired urinary tract infections in patients with diabetes and urinary catheterization. J Exp Clin Med. 2014;6:90–93.

[30] Aloush SM, Al Qadire M, Assmairan K, et al. Risk factors for hospital-acquired non-catheter-associated urinary tract infection. J Am Assoc Nurse Pract. 2019;31:747–751.

[31] Bearman GML, Wenzel RP. Bacteremias: A leading cause of death. Arch Med Res. 2005;36:646–659.

[32] Mortensen VH, Søgaard M, Mygind LH, et al. Incidence and mortality of hospital-acquired bacteraemia: a population-based cohort study applying a multi-state model approach. Clin Microbiol Infect. 2022.

[33] ECDC, EMEA. The bacterial challenge: time to react. 2009.

[34] Conway LJ, Carter EJ, Larson EL. Risk Factors for Nosocomial Bacteremia Secondary to Urinary Catheter-Associated Bacteriuria: A Systematic Review. Urol Nurs. 2015;35:191–203.

[35] Munoz-Price LS, Dezfulian C, Wyckoff M, et al. Effectiveness of stepwise interventions targeted to decrease central catheter-associated bloodstream infections. Crit Care Med. 2012;40:1464–1469.

[36] Riu M, Chiarello P, Terradas R, et al. Cost attributable to nosocomial bacteremia. Analysis according to microorganism and antimicrobial sensitivity in a university hospital in Barcelona. PLoS One. 2016;11:1–12.

[37] Mitchell BG, Ferguson JK, Anderson M, et al. Length of stay and mortality associated with healthcare-associated urinary tract infections: A multi-state model. J Hosp Infect. 2016;93:92–99.

[38] Saint S. Clinical and economic consequences of nosocomial catheter-related bacteriuria. Am J Infect Control. 2000;28:68–75.

[39] Pirson M, Dramaix M, Struelens M, et al. Costs associated with hospital-acquired bacteraemia in a Belgian hospital. J Hosp Infect. 2005;59:33–40.

[40] Markovic-Denic L, Mijovic B, Jankovic S. Risk factors for hospital-acquired urinary tract infection: A case-control study. Int Urol Nephrol. 2011;43:303–308.

[41] King C, Garcia Alvarez L, Holmes A, et al. Risk factors for healthcare-associated urinary tract infection and their applications in surveillance using hospital administrative data: A systematic review. J Hosp Infect. 2012;82:219–226.

[42] Keren R, Shaikh N, Pohl H, et al. Risk Factors for Recurrent Urinary Tract Infection and Renal Scarring. Pediatrics. 2015;136:e13–e21.

[43] Redder JD, Leth RA, Møller JK. Analysing risk factors for urinary tract infection based on automated monitoring of hospital-acquired infection. J Hosp Infect. 2016;92:397–400.

[44] Foxman B. Urinary tract infection syndromes. Occurrence, recurrence, bacteriology, risk factors, and disease burden. Infect Dis Clin North Am. 2014;28:1–13.

[45] Mortensen VH, Søgaard M, Mygind LH, et al. Incidence and mortality of hospital-acquired bacteraemia: a population-based cohort study applying a multi-state model approach. Clin Microbiol Infect. 2022;1–7.

[46] Schønheyder HC, Søgaard M. Hospitalserhvervet bakteriæmi og fungæmi. Ugeskr Laeger. 2007;169:4175–4179.

[47] Butler-Laporte G, Cheng MP, McDonald EG, et al. Screening swabs surpass traditional risk factors as predictors of MRSA bacteremia. BMC Infect Dis. 2018;18:1–6.

[48] Dua S, Rajendra U, Prerna A, et al. Machine Learning in Healthcare Informatics. 2014.

[49] Scardoni A, Balzarini F, Signorelli C, et al. Artificial intelligence-based tools to control healthcare associated infections: A systematic review of the literature. J Infect Public Health. 2020;13:1061–1077.

[50] Gubbels S, Nielsen J, Voldstedlund M, et al. National automated surveillance of hospital-acquired bacteremia in Denmark using a computer algorithm. Infect Control Hosp Epidemiol. 2017;38:559–566.

[51] Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018;2:719–731.

[52] Randall Moorman J. The principles of whole-hospital predictive analytics monitoring for clinical medicine originated in the neonatal ICU. npj Digit Med. 2022;5:1–6.

[53] Duda RO, Hart PE. Pattern Classification. 2001.

[54] Topol EJ. Deep medicine: how artificial intelligence can make healthcare human again. First edit. Basic Books, New York; 2019.

[55] Thorsen-Meyer HC, Nielsen AB, Nielsen AP, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. Lancet Digit Heal. 2020;2:e179–e191.

[56] Peiffer-Smadja N, Rawson TM, Ahmad R, et al. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. Clin Microbiol Infect. 2019.

[57] Luz CF, Vollmer M, Decruyenaere J, et al. Machine learning in infection management using routine electronic health records: tools, techniques, and reporting of future technologies. Clin Microbiol Infect. 2020.

[58] Tang R, Luo R, Tang S, et al. Machine learning in predicting antimicrobial resistance: a systematic review and meta-analysis. Int J Antimicrob Agents. 2022;60.

[59] Lipton ZC. The mythos of model interpretability. Work Hum Interpret Mach Learn (WHI 2016). 2016;61:35–43.

[60] Burkart N, Huber MF. A survey on the explainability of supervised machine learning. J Artif Intell Res.

2020;70:245–317.

[61] Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. NAACL-HLT 2016 - 2016 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol Proc Demonstr Sess. 2016;97–101.

[62] Derks IP, de Waal A. A Taxonomy of Explainable Bayesian Networks. Commun Comput Inf Sci. 2020;1342:220–235.

[63] Vellido A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Comput Appl. 2020;32:18069–18083.

[64] Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019;25.

[65] Verma S, Dickerson J, Hines K. Counterfactual Explanations for Machine Learning: Challenges Revisited. 2021.

[66] Jeng SL, Huang ZJ, Yang DC, et al. Machine learning to predict the development of recurrent urinary tract infection related to single uropathogen, Escherichia coli. Sci Rep. 2022;12:1–11.

[67] Møller JK, Sørensen M, Hardahl C. Prediction of risk of acquiring urinary tract infection during hospital stay based on machine-learning: A retrospective cohort study. PLoS One. 2021;16:1–16.

[68] Zhu C, Xu Z, Gu Y, et al. Prediction of Poststroke Urinary Tract Infection Risk in Immobile Patients Using Machine Learning: a observational cohort study. J Hosp Infect. 2022.

[69] Taylor RA, Moore CL, Cheung KH, et al. Predicting urinary tract infections in the emergency department with machine learning. PLoS One. 2018;13:1–15.

[70] Yelin I, Snitser O, Novich G, et al. Personal clinical history predicts antibiotic resistance of urinary tract infections. Nat Med. 2019;25:1143–1152.

[71] Park JI. Developing a Predictive Model for Hospital-Acquired Catheter-Associated Urinary Tract Infections Using Electronic Health Records and Nurse Staffing Data. 2016.

[72] Garnica O, Gómez D, Ramos V, et al. Diagnosing hospital bacteraemia in the framework of predictive, preventive and personalised medicine using electronic health records and machine learning classifiers. EPMA J. 2021;12:365–381.

[73] Choi DH, Hong KJ, Park JH, et al. Prediction of bacteremia at the emergency department during triage and disposition stages using machine learning models. Am J Emerg Med. 2022.

[74] Ratzinger F, Dedeyan M, Rammerstorfer M, et al. A risk prediction model for screening bacteremic patients: A cross sectional study. PLoS One. 2014;9.

[75] Ward L, Møller JK, Eliakim-Raz N, et al. Prediction of Bacteraemia and of 30-day Mortality Among Patients with Suspected Infection using a CPN Model of Systemic Inflammation. IFAC-PapersOnLine. 2018.

[76] Mahmoud E, Dhoayan M Al, Bosaeed M, et al. Developing machine-learning prediction algorithm for bacteremia in admitted patients. Infect Drug Resist. 2021;14:757–765.

[77] Ratzinger F, Haslacher H, Perkmann T, et al. Machine learning for fast identification of bacteraemia in SIRS patients treated on standard care wards: a cohort study. Sci Rep. 2018.

[78] Caruana R, Lou Y, Gehrke J, et al. Intelligible Models for HealthCare. 2015;1721–1730.

[79] Jialiang Jiang. Interpretable Learning for Hospital Readmission Prediction from Healthcare Data. 2020.

[80] Shaban-Nejad A, Michalowski M, Buckeridge DL, et al. Explainable AI in Healthcare and Medicine. 2020.

[81] Tjoa E, Guan C. A Survey on Explainable Artificial Intelligence. 2015;14:1–22.

[82] Tjoa E, Guan C. A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. IEEE Trans Neural Networks Learn Syst. 2021;32:4793–4813.

[83] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206–215.

[84] Stiglic G, Kocbek P, Fijacko N, et al. Interpretability of machine learning-based prediction models in healthcare. Wiley Interdiscip Rev Data Min Knowl Discov. 2020;10:1–13.

[85] Holzinger A, Langs G, Denk H, et al. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov. 2019;9:1–13.

[86] Ahmad MA, Teredesai A, Eckert C. Interpretable machine learning in healthcare. Proc - 2018 IEEE Int Conf Healthc Informatics, ICHI 2018. 2018;447.

[87] Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning. 2017;1–13.

[88] Molnar C, Casalicchio G, Bischl B. Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. Commun Comput Inf Sci. 2020;1323:417–431.

[89] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;2017-Decem:4766–4775.

[90] Wachter S, Mittelstadt B, Russell C. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. SSRN Electron J. 2017;1–52.

[91] Stenwig E, Salvi G, Rossi PS, et al. Comparative analysis of explainable machine learning prediction models for hospital mortality. 2022;1–14.

[92] Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58:82–115.

[93] Gilpin LH, Bau D, Yuan BZ, et al. Explaining Explanations : An Overview of Interpretability of Machine Learning. 2019;

[94] Gunning D, Aha DW. DARPA's explainable artificial intelligence program. AI Mag. 2019;40:44–58.

[95] McCoy LG, Nagaraj S, Morgado F, et al. What do medical students actually need to know about artificial intelligence? npj Digit Med. 2020.

[96] Laxminarayan R, Duse A, Wattal C, et al. Antibiotic resistance-the need for global solutions. Lancet Infect Dis. 2013;13:1057–1098.

[97] O'Neill J. Antimicrobial Resistance: Tackling a crisis for the health and wealth of nations The. 2014.

[98] Henius AE, Pedersen K, Jensen LB. Danmap 2017. 2017;

[99] Ahmad M, Khan AU. Global economic impact of antibiotic resistance: A review. J Glob Antimicrob Resist. 2019;19:313–316.

[100] Sundararajan V, Henderson T, Perry C, et al. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. J Clin Epidemiol. 2004;57:1288–1294.

[101] Hansen BV. Acute admissions to internal medicine departments in Denmark - studies on admission rate , diagnosis , and prognosis - PhD Dissertation. 2014;

[102] Christensen DH, Veres K, Ording AG, et al. Risk of cancer in patients with thyroid disease and venous thromboembolism. Clin Epidemiol. 2018;10:907–915.

[103] Nielsen PR, Benros ME, Mortensen PB. Hospital contacts with infection and risk of schizophrenia: A population-based cohort study with linkage of danish national registers. Schizophr Bull. 2014;40:1526–1532.

[104] Rapsang AG, Shyam DC. Scoring systems in the intensive care unit: A compendium. Indian J Crit Care Med. 2014;18:220.

[105] Zhang Q, Shi B, Wu L. Characteristics and risk factors of urinary tract infection in patients with HBV-related acute-on-chronic liver failure: A retrospective study. Med (United States). 2022;101:E29913.

[106] Yoshida B, Nguyen A, Formanek B, et al. Hypoalbuminemia and Elevated CRP are Risk Factors for Deep Infections and Urinary Tract Infections After Lumbar Spine Surgery in a Large Retrospective Patient Population. Glob Spine J. 2021;13:33–44.

[107] Jensen F V, Nielsen TD. Bayesian networks and decision graphs. Inf. Sci. Stat. 2007.

[108] Lauritzen SL. The EM algorithm for graphical association models with missing data. Comput Stat Data Anal. 1995;19:191–201.

[109] Gupta A, Liu T, Shepherd S. Clinical decision support system to assess the risk of sepsis using Tree Augmented Bayesian networks and electronic medical record data. Health Informatics J. 2020;26:841–861.

[110] Friedman N, Geiger D, Goldzmidt M. Bayesian network classifiers. 1997;

[111] Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search. 2000.

[112] Le TD, Hoang T, Li J, et al. A fast PC algorithm for high dimensional causal discovery with multi-core PCs. IEEE/ACM Trans Comput Biol Bioinforma. 2019;16:1483–1495.

[113] Iguyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157–1182.

[114] Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. London, Edinburgh, Dublin Philos Mag J Sci. 1900;50:157–175.

[115] Pedregosa F, Varoquaux G, Thirion AGVMB. Scikit-learn: Machine Learning in Python. J ofMachine Learn Res. 2011;12:2825–2830.

[116] Breiman L. Classification and regression trees. Breiman L, editor. Boca Raton, Fla: Chapman & Hall; 1984.

[117] Lemaıtre G, Nogueira F, Aridas CK. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. J Mach Learn Res. 2017;18:1–5.

[118] Laupacis R, Sackett A, Roberts D. An assessment of clinically useful measures of the consequences of treatment. 1988;318:1728–1733.

[119] Rickard CM, Webster J, Wallis MC, et al. Routine versus clinically indicated replacement of peripheral

intravenous catheters: A randomised controlled equivalence trial. Lancet. 2012;380:1066–1074.

[120] Meirson T, Goldman A, Bomze D. For and Against Routine Removal of Peripheral Intravenous Catheters. JAMA Intern Med. 2022;455–456.

[121] Rautenberg T, Gerritsen A, Downes M. Health economic decision tree models of diagnostics for dummies: A pictorial primer. Diagnostics. 2020;10.

[122] Drummond MF, Sculpher MJ, Claxton K, et al. Methods for the Economic Evaluation of Health Care Programmes. Oxford: Oxford University Press; 2015.

[123] Finansministeriet. Dokumentationsnotat – den samfundsøkonomiske diskonteringsrente. 2021;1–19.

[124] Behandlingsraadet. Omkostningsopgørelse. 2022;1–36.

[125] Buetti N, Abbas M, Pittet D, et al. Comparison of Routine Replacement with Clinically Indicated Replacement of Peripheral Intravenous Catheters. JAMA Intern Med. 2021;181:1471–1478.

[126] Lacave C, Díez FJ. A review of Explanation Methods for Bayesian Network. 2002;

[127] Lacave C, Díez FJ. A review of explanation methods for Bayesian networks. Knowl Eng Rev. 2002;17:107–127.

[128] Sejer Jakobsen R, Hejlesen O, Cichosz SL, et al. Developing a Bayesian Network as a decision support system for evaluating patients with diabetes mellitus admitted to the intensive care unit - a proof of concept. 2018;

[129] Cooper GF, Herskovits E. A Bayesian Method for the Induction of Probabilistic Networks from Data. Mach Learn. 1992;9:309–347.

[130] Knaus, William A. M, Draper, Elizabeth A., MS Wagner, Douglas P. PhD, Zimmerman JEM. APACHE II: A severity of disease classification system. Crit Care Med. 1985;13:818–829.

[131] Vilone G, Longo L. Explainable Artificial Intelligence: a Systematic Review. 2020.

[132] Wiedermann CJ. Hypoalbuminemia as surrogate and culprit of infections. Int J Mol Sci. 2021;22.

[133] Storme O, Saucedo JT, Garcia-Mora A, et al. Risk factors and predisposing conditions for urinary tract infection. Ther Adv Urol. 2019;11:19–28.

[134] Scherberich JE, Fünfstück R, Naber KG. Urinary tract infections in patients with renal insufficiency and dialysis - epidemiology, pathogenesis, clinical symptoms, diagnosis and treatment. GMS Infect Dis. 2021;9:Doc07.

[135] Delgado A, Cordero G-G E, Marcos S, et al. Influence of cognitive impairment on mortality, complications and functional outcome after hip fracture: Dementia as a risk factor for sepsis and urinary infection. Injury. 2020;51:S19–S24.

[136] Wang MC, Tseng CC, Wu AB, et al. Bacterial characteristics and glycemic control in diabetic patients with Escherichia coli urinary tract infection. J Microbiol Immunol Infect. 2013;46:24–29.

[137] Deshmukh F, Merchant SS. Explainable Machine Learning Model for Predicting GI Bleed Mortality in the Intensive Care Unit. Am J Gastroenterol. 2020;115:1657–1668.

[138] Lauritsen SM, Kristensen M, Olsen MV, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat Commun. 2020;11:1–11.

[139] Parreco JP, Hidalgo AE, Badilla AD, et al. Predicting central line-associated bloodstream infections and mortality

using supervised machine learning. J Crit Care. 2018;45:156–162.

[140] Pollard TJ, Johnson AEW, Raffa JD, et al. The eICU collaborative research database, a freely available multi-center database for critical care research. Sci Data. 2018;5:1–13.

[141] Beeler C, Dbeibo L, Kelley K, et al. Assessing patient risk of central line-associated bacteremia via machine learning. Am J Infect Control. 2018;46:986–991.

[142] Tuffaha, HW, Rickard, CM, Webster, J, Marsh, N, Gordon, L, Wallis, M, Scuffham P. Cost-effectiveness analysis of clinically-indicated versus routine replacement of peripheral intravenous catheters. Appl Health Econ Health Policy. 2014;12:51–58.

[143] Webster J, Osborne S, Rickard CM NK, Webster. Clinically-indicated replacement versus routine replacement of peripheral venous catheters. 2013;39:67–69.

[144] Zingg W, Barton A, Bitmead J, et al. Best practice in the use of peripheral venous catheters: A scoping review and expert consensus. Infect Prev Pract. 2023;5:100271.

[145] Webster J, Osborne S, Rickard CM MN. Clinically-indicated replacement versus routine replacement of peripheral venous catheters (Review). 2019;327–377.

[146] Chen CY, Chen WC, Chen JY, et al. Comparison of clinically indicated replacement and routine replacement of peripheral intravenous catheters: A systematic review and meta-analysis of randomized controlled trials. Front Med. 2022;9:1–10.

[147] Roberts RR, Hota B, Ahmad I, et al. Hospital and Societal Costs of Antimicrobial-Resistant Infections in a Chicago Teaching Hospital: Implications for Antibiotic Stewardship. Clin Infect Dis. 2009;49:1175–1184.

[148] Bysshe T, Gao Y, Al. H-HK et. Estimating the Additional Hospital Inpatient Cost and Mortality Associated With Selected Hospital-Acquired Conditions | Agency for Healthcare Research and Quality. Natl. Score Card Reports. 2017.

[149] Olesen B, Anhøj J, Rasmussen KP, et al. MiBAlert—a new information tool to fight multidrug-resistant bacteria in the hospital setting. Int J Med Inform. 2016;95:43–48.

[150] Mori T, Uchihira N. Balancing the trade-off between accuracy and interpretability in software defect prediction. Empir. Softw. Eng. Empirical Software Engineering; 2019.

[151] Xu F, Uszkoreit H, Du Y, et al. Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. 2019.

[152] Su CH, Chang SC, Yan JJ, et al. Excess Mortality and Long-Term Disability from Healthcare-Associated Staphylococcus aureus Infections: A Population-Based Matched Cohort Study. PLoS One. 2013;8.

[153] Sejer Jakobsen H, Brix J, Sejer Jakobsen R. Unraveling Data From an Idea Management System of 11 Radical Innovation Portfolios Key Lessons and Avenues for Articial Intelligence Integration. 2023;30.

[154] Thomas J. Force Field Analysis : A New Way to Evaluate Your Strategy. 1985;18:54–59.

PAPERS

STUDY I: A STUDY ON THE RISK STRATIFICATION FOR PATIENTS WITHIN 24 HOURS OF ADMISSION FOR RISK OF HOSPITAL-ACQUIRED URINARY TRACT INFECTION USING EXPLAINABLE BAYESIAN NETWORK MODELS

Rune Sejer Jakobsen[1, 2], Thomas Dyhre Nielsen[3], Peter Leutscher [1, 4] and Kristoffer Koch[1, 5]

a) Centre for Clinical Research, North Denmark Regional Hospital, Hjørring, Denmark
b) Business Intelligence and Analysis, The North Denmark Region, Denmark
c) Department of Computer Science, Aalborg University, Denmark
d) Department of Clinical Medicine, Aalborg University, Denmark
e) Department of Clinical Microbiology, Aalborg University Hospital, Denmark

The paper is submitted and has been made available for the Ph.D. committee.

Not available in the published version of the dissertation.

STUDY II: CLINICAL EXPLAINABLE MACHINE LEARNING MODELS FOR EARLY IDENTIFICATION OF PATIENTS AT RISK OF HOSPITAL-ACQUIRED URINARY TRACT INFECTION

Journal Pre-proof

Clinical explainable machine learning models for early identification of patients at risk of hospital-acquired urinary tract infection

Rune Sejer Jakobsen, Thomas Dyhre Nielsen, Peter Leutscher, Kristoffer Koch

PII: S0195-6701(23)00106-8

DOI: https://doi.org/10.1016/j.jhin.2023.03.017

Reference: YJHIN 6894

To appear in: *Journal of Hospital Infection*

Received Date: 21 February 2023

Accepted Date: 22 March 2023

Please cite this article as: Jakobsen RS, Nielsen TD, Leutscher P, Koch K, Clinical explainable machine learning models for early identification of patients at risk of hospital-acquired urinary tract infection, *Journal of Hospital Infection*, https://doi.org/10.1016/j.jhin.2023.03.017.

Clinical explainable machine learning models for early identification of patients at risk of hospital-acquired urinary tract infection

Rune Sejer Jakobsen[a, b], Thomas Dyhre Nielsen[c], Peter Leutscher [a, d] and Kristoffer Koch[a, e]

f) Centre for Clinical Research, North Denmark Regional Hospital, Hjørring, Denmark

g) Business Intelligence and Analysis, The North Denmark Region, Denmark

h) Department of Computer Science, Aalborg University, Denmark

i) Department of Clinical Medicine, Aalborg University, Denmark

j) Department of Clinical Microbiology, Aalborg University Hospital, Denmark

Keywords

Clinical explainability, feature selection, hospital-acquired urinary tract infection, machine learning, risk prediction

Summary

Background: Machine learning (ML) models for early identification of patients at risk of hospital-acquired urinary tract infections (HA-UTI) may enable timely and targeted preventive and therapeutic strategies. However, clinicians are often challenged in the interpretation of the predictive outcomes provided by the ML models, which often reach different performances.

Aim: To train ML models for predicting patients at risk of HA-UTI using available data from electronic health records at the time of hospital admission. We focused on the performance of different ML models and clinical explainability.

Methods: This retrospective study investigated patient data representing 138.560 hospital admissions in the North Denmark Region from 01.01.2017 to 31.12.2018. We extracted 51 health socio-demographic and clinical features in a full dataset and used the χ^2 test in addition to expert knowledge for feature selection, resulting in two reduced datasets. Seven different ML models were trained and compared between the three datasets. We applied the SHapley Additive exPlanation (SHAP) method to support population- and patient-level explainability.

Findings: The best-performing ML model was a neural network based on the full dataset, reaching an area under the curve (AUC) of 0.758. The neural network was also the best-performing ML model based on the reduced datasets, reaching an AUC of 0.746. Clinical explainability was demonstrated with a SHAP summary- and forceplot.

Conclusion: Within 24h of hospital admission, the ML models were able to identify patients at risk of developing HA-UTI, providing new opportunities to develop efficient strategies for the prevention of HA-UTI. Using SHAP, we demonstrate how risk predictions can be explained at individual patient level and for the patient population in general.

Introduction

Hospital-acquired urinary tract infection (HA-UTI) is a common yet often preventable complication during hospital admission [1]. HA-UTI can lead to prolonged length of stay [2], has an attributable mortality estimated to be 9–13% [3,4], and is associated with increased hospital expenditures [2]. Different strategies for control and management of HA-UTI have been widely implemented [5], e.g., preventive hygienic measures [6], urinary catheter care bundles [7], and rational use of antibiotics [8]. However, the incidence of HA-UTI remains continuously high, and the negative consequences of HA-UTI will likely increase in the coming years due to the emergence of multi-resistant urinary tract pathogens [9,10]. The need for new tools in the control of HA-UTI is obvious.

With the digitalization of the healthcare system, machine learning (ML) models provide new opportunities to develop efficient strategies for the prevention of Hospital-associated infections, e.g., aiming at early prediction of patients at risk of HA-UTI [9,11]. However, clinicians may lack experience, confidence, and trust in the predictions of these models [12,13]. This situation may negatively impact the adoption into the clinical routine. Future development of clinical ML models should strive to include clinical expert knowledge in the overall development process and provide an explanation of the ML models' clinical prediction [12–17].

Integration of clinical explainability in relation to the development of ML models in the healthcare sector is a many-faced challenge traditionally categorized into model-specific or model-agnostic techniques, which can be further categorized as population-level (global) and patient-level (local) explainability [16,18]. Model-specific explainability techniques refer to ante-hoc evaluations in specific classes of ML models, which may be performed for specific ML model classes, e.g., logistic regression (LR), Bayesian network (BN), and decision trees (DT), where the inner mechanics of the model can be examined by, e.g., inspecting internal model parameters, dependencies, the rules for the splits in a tree, etc. [16,18]. On the other hand, model-agnostic explainability refers to a post hoc evaluation of the predictions of a given ML model [16,19], enabling support for explainability of any ML model class, including black-box neural network (NN) or ensemble models, such as random forest (RF), Adaboost (AD), and gradient boosting (GB). In this scenario, Stenwig et al. [13] and Deshmukh et al. [20] recently demonstrated the use of the model-agnostic SHapley Additive exPlanation (SHAP)-method [21] and its applicability to a clinical dataset [22]. The SHAP method calculates Shapley values (known from game theory) and visualizes the contribution of each feature to the prediction of a given target, supporting reasoning about the predictions obtained from an ML model. To the best of our knowledge, no previous work has compared performances and evaluated the explainability of ML models for the prediction of patients at risk of HA-UTI using the model-agnostic SHAP method.

We aim to train, test, and compare different ML models and demonstrate how the SHAP method can support both population- and patient-level explainability in the risk prediction of HA-UTI. Moreover, in the training of our explainable ML models, we identify the most clinically meaningful features for predicting HA-UTI.

Methods

Study design

This retrospective study is conducted in collaboration between the Centre for Clinical Research at the North Denmark Regional Hospital, the Department of Computer Science at the University of Aalborg in Denmark, and the Business Intelligence and Analysis Unit in the North Denmark Region. Danish legislation requires no consent from participants or approval from an ethics committee for registry-based studies.

Participants

We include data from electronic patient health records on all hospitalized adult persons age >18 years with a hospital admission between 1 January 2017 and 31 December 2018 at North Denmark Regional Hospital or Aalborg University Hospital.

Hospital-acquired urinary tract infections

Statens Serum Institut in Denmark provide data on HA-UTI from their national surveillance system, called the Hospital-Acquired Infections Database (HAIBA) [23,24]. The definition of HA-UTI in HAIBA is based on a combination of the following information: urine culture findings, a diagnosis code indicating UTI, and a relevant course of antibiotic treatment. The case definition has previously been described in detail by Condell et al. [23]. HA-UTI is included as a binary target feature in our dataset.

Model design

The ML models seek to enable stratification of patients within the first 24h of admission according to the risk of developing HA-UTI. The approach can be considered a warning system for the identification of patients who may benefit from early-onset preventive measures and, thereby, a tool for guided awareness, e.g., deciding which patients need targeted urinary catheter care bundles to prevent HA-UTI.

Predictive features

We grade the severity of each feature in discrete levels using clinical expert knowledge and include only the worst-graded measure within 24h for each predictive feature of each admission in our dataset. Supplementary A presents our complete set of 51 health socio-demographic and clinical features and their graded discrete levels.

Supplementary A: Discretization of the predictive features

Feature	Discrete level 0	Discrete level 1	Discrete level 2	Discrete level 3	Discrete level 4	Discrete level 5	Discrete level 6	unit
Admission detail								
Triage	Blue	Green	Yellow	Orange	Red	-	-	-
Season	Winter	Fall	Spring	Summer	-	-	-	-
Health socio-demographics								
Age	18 – 29	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	> 80	Years
Gender	Female	Male	-	-	-	-	-	-
Body Mass Index	< 18.49	18.5 – 24.9	25. – 29.9	30 – 34.9	35. – 39.9	> 40	-	kg/m^2
Smoking	No	Ex-smoker	Yes	-	-	-	-	-
Alcoholic	No	Yes	-	-	-	-	-	> 14w and
Exercise	No	Yes	-	-	-	-	-	>21m[15]
Comorbidities								> 30min/d[16]
Number of comorbidities	0	1	2	> 2	-	-	-	
Acute myocard.	No	Yes	-	-	-	-	-	-
infarction	No	Yes	-	-	-	-	-	-
Congestive heart failure	No	Yes	-	-	-	-	-	-
Peripheral vascular	No	Yes	-	-	-	-	-	-
disease	No	Yes	-	-	-	-	-	-
Cerebral vascular	No	Yes	-	-	-	-	-	-
accident	No	Yes	-	-	-	-	-	-
Dementia	No	Yes	-	-	-	-	-	-
Pulmonary disease	No	Yes	-	-	-	-	-	-
Connective tissue	No	Yes	-	-	-	-	-	-
disorder	No	Yes	-	-	-	-	-	-
Peptic ulcer	No	Yes	-	-	-	-	-	-
Liver disease	No	Yes	-	-	-	-	-	-
Diabetes	No	Yes	-	-	-	-	-	-
Diabetes complications	No	Yes	-	-	-	-	-	-
Paraplegia	No	Yes	-	-	-	-	-	-
Renal disease	No	Yes	-	-	-	-	-	-
Cancer	No	Yes	-	-	-	-	-	-
Metastatic cancer	No	Yes	-					-

[15] The limit of being alcoholic is set to >14 drinks a week for women and > 21 drinks a week for men.

[16] The limit for exercising sufficiently is set to > 30 minutes per day.

Leukemia								-
Lymphoma								
Severe liver disease								
History of urinary tract infection								
Vital parameters								
Temperature	36 – 37.9	35. – 35.9	38. – 38.9	39 – 39.9	> 40.	< 35.	-	^{O} Celsius
Glasgow Coma scale	15	14	9 – 13	3 – 8	-	-	-	-
Respiratory rate	12 – 20	8 – 11	21 – 24	25 – 29	30 – 35	> 36	< 8	Breaths pm
Pulse	50 – 89	90 – 109	40 – 49	110 – 129	> 130	< 40	-	Beats pm
Systolic blood pressure	110 – 219	100 – 110	90 – 99	> 220	< 90	-	-	mmHg
Diastolic blood pressure	60 – 89	90 – 99	100 – 119	> 120	< 60	-	-	mmHg
Oxygen saturation	> 96	94 – 95	92 – 93	80 – 91	< 80	-	-	%
Laboratory results								
B-thrombocytes	145 – 389.9	60 – 139.9	390 – 999.9	20 – 59.9	< 19.9	> 1000	-	x 10^{9}/L
B-erythrocytes	3.9 – 5.69	5.7 – 7.49	> 7.5	2.5 – 3.89	< 2.49	-	-	x 10^{12}/L
P- partial pressure of oxygen	> 9.33	8. – 9.32	7.33 – 7.99	< 7.32	-	-	-	kPa
	< 7.99	8. – 19.99	20. – 39.9	40. – 199.9	> 200.	-	-	Mg/L
P- C-reactive protein	7.3 – 10.49	4. – 7.29	10.5 – 12.49	> 12.5	< 3.99	-	-	mmol/L
B-hemoglobin	34 – 47.99	> 48.	16 – 33.99	> 16	-	-	-	g/L
P-albumin	45 – 104.99	20 – 44.99	105 – 119.99	< 19.99	> 120	-	-	µmol/L
P-creatinine	5 – 24.99	< 4.99	25 – 59.99	60 – 79.99	> 80	-	-	µmol/L
P-bilirubin	4 – 7.99	8 – 11.99	12 – 19.99	20 – 32.99	2 – 3.99	> 33	< 2	mmol/L
P-glucose	7.33 – 7.49	7.5 – 7.59	7.25 – 7.32	7.6 – 7.69	7.15 – 7.24	< 7.14	> 7.7	-
P-pH	0.5 – 2.49	< 0.49	2.5 – 8.99	> 9.	-	-	-	mmol/L
P-lactate	3.5 – 8.79	8.8 – 19.99	20 – 49.99	0.5 – 3.49	> 50	< 0.49	-	x 10^{9}/L
B-leucocytes	2 – 8.79	0.5 – 1.99	8.8 – 19.99	20 – 49.99	> 50	< 0.49	-	x 10^{9}/L
B-neutrofilocytes	< 0.09	0.1 – 0.19	0.2 – 0.29	0.3 – 0.39	> 0.4	-	-	x 10^{9}/L
B-monocytes								
Procedure	No	Yes	-	-	-	-	-	-
Urinary catheter								

The International Classification of Diseases, version 10 (ICD-10), is used for categorizing admission causes, where a primary diagnosis is registered for each admission. We exclude *External causes of morbidity and mortality* because it is not used for reporting primary diagnosis, in addition to *Certain conditions originating in the perinatal period*, as patients in our dataset were > 18 years of age. Supplementary B presents the 19 admission causes and their associated ICD-10 codes for the admission cause feature.

Supplementary B: The codes for the included admission causes

#	Admission cause	ICD-10 codes
1	Certain infectious and parasitic diseases	A00 – A99, B00 – B99
2	Neoplasms	C00 – C99, D00 – D49
3	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	D50 – D99
4	Endocrine, nutritional and metabolic diseases	E00 – E99
5	Mental and behavioral disorders	F00 – F99
6	Diseases of the nervous system	G00 –G99
7	Diseases of the eye and adnexa	H00 – H59
8	Diseases of the ear and mastoid process	H60 – H95
9	Diseases of the circulatory system	I00 – I99
10	Diseases of the respiratory system	J00 – J99
11	Diseases of the digestive system	K00 – K93
12	Diseases of the skin and subcutaneous tissue	L00 – L99
13	Diseases of the musculoskeletal system and connective tissue	M00 – M99
14	Diseases of the genitourinary system	N00 – N99
15	Pregnancy, childbirth and the puerperium	O00 – O99
16	Certain conditions originating in the perinatal period	P00 – P96
17	Congenital malformations, deformations, and chromosomal abnormalities	Q00 – Q99
18	Symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified	R00 – R99
19	Injury, poisoning, and certain other consequences of external causes	S00 – S99, T00 – T98
20	External causes of morbidity and mortality	X60 – X98, Y00 – Y09
21	Factors influencing health status and contact with health services	Z00 – Z99

ICD-10 and version 8 (ICD-8), respectively, are used for comorbidities according to the Charlson Comorbidities Index (CCI). The selection of relevant ICD-10 codes for CCI is inspired by the work of Sundararajan et al. [25], Hansen et al. [26], and Christensen et al. [27], while the selection of ICD-8 codes for CCI is influenced by the research of Christensen et al. [27]. The ICD-10 codes for a history of urinary tract infection (UTI) are inspired by the studies of Nielsen et al. [28] and Gubbels et al. [24], while the ICD-8 codes for a history of UTI are influenced by the research of Nielsen et al. [28]. However, one predictor, human immunodeficiency virus (HIV), with a frequency of less than 50 occurrences in the dataset, is excluded from the model as there are no HA-UTI cases in the study population with this comorbidity. Supplementary C presents the selected comorbidities and their associated ICD-10 and ICD-8 codes.

Supplementary C: The codes for the included comorbidities

Comorbidities

#	features	ICD-10 codes	ICD-8 codes
1	Acute myocardial infarction	I21 – I23, I25	410
2	Congestive heart failure	I50, I11, I13	042709 – 42711, 42719, 42899, 78249
3	Peripheral vascular disease	R02, Z95, I70, I71 – I79	440 – 445
4	Cerebral vascular accident	I60 –I67, G45, G46	430 – 438
5	Dementia	F00 –F03, F05, G30	29009 – 29019, 29309
6	Pulmonary disease	J00 – J86, J90 – J99	490 – 493, 515 – 518
7	Connective tissue disorder	M05 – M14, M30 – M36, D86	13599, 4446, 712, 716, 734
8	Peptic ulcer	K22, K25, K26, K27, K28	53091, 53098, 531, 532, 533, 534
9	Liver disease	B17, B18, K71, K73, K74, K75, K77	571, 57301, 57304
10	Diabetes	E10, E100, E101, E109, E109A, E11, E110, E111, E119,	24900, 24906, 24907, 24909, 25000, 25006, 25007, 25009
	-	E13, E130, E131, E135 – E139, E14, E140 – E149	-
11	Diabetes complications	E102 – E105, E105A, E105B, E105C, E105D, E106 –	24901 – 24905, 24908, 25001 – 25005, 25008
	-	E108, E112 – E118, E132, E133, E134, E142 – E148	-
12	Paraplegia	DG04, DG80, DG81, DG82, DG83	344
13	Renal disease	N01, N02, N03, N04, N05, N06, N07, N08	403, 404, 580 – 584, 59009, 59319, 75310 – 75318, 792
14	Cancer	C00 – C76, C86	140 – 194, 19248
15	Metastatic cancer	C77 – C80	195 – 199
16	Leukemia	C91 – C95	204 – 207
17	Lymphoma	C81 – C85, C88, C90, C96	200 – 203, 27559
18	Severe liver disease	B15, B16, B19, K70, K72, K76, I85	07000, 07002, 07004, 07006, 07008, 45600 –45609, 57300
	-	-	07983
19	HIV	B20 – B24	580, 590, 59500, 59501
20	History of UTI	N30, N390, N390, O234, P393, O233, O239, O862, T814U	-

To reduce the noise from potentially less meaningful features and achieve faster, more robust, and more efficient ML models [29], we perform an automated feature selection based on automated χ^2 [30] and a manual feature selection based on expert knowledge. The scikit-learn (sklearn) machine learning environment version 1.0.2 (https://scikit-learn.org/) [31] is used for the automated feature selection. We use suggestions on risk factors for HA-UTI from the literature [32–36] and suggestions from clinical experts for the expert knowledge dataset, where at least one feature within each of the following categories is represented: health socio-demographics, comorbidities, vital parameters, laboratory results, and urinary catheter usage. After the feature selection, the two reduced datasets each include ten features. Table 1 presents the selected features for our full-, χ^2-, and expert knowledge datasets.

Table 1: Predictive features included in each dataset for predicting the risk of hospital-acquired urinary tract infection.

Features	Full	χ^2	Expert knowledge
Admission details			
Admission cause	✓	✓	
Triage	✓		
Season	✓		
Health socio-demographics			
Age	✓	✓	✓
Gender	✓		✓
Body Mass Index (BMI)	✓		
Smoking	✓		
Alcohol	✓		
Exercise	✓		
Comorbidities			
Number of comorbidities	✓	✓	
Acute myocardial infarction	✓		
Congestive heart failure	✓		
Peripheral vascular disease	✓		
Cerebral vascular accident	✓		
Dementia	✓		
Pulmonary disease	✓		
Connective tissue disorder	✓		
Peptic ulcer	✓		
Liver disease	✓		
Diabetes	✓		✓
Diabetes complications	✓		
Paraplegia	✓		
Renal disease	✓		
Cancer	✓		
Metastatic cancer	✓		
Leukemia	✓		
Lymphoma	✓		
Severe liver disease	✓		
History of urinary tract infection	✓		✓
Vital parameters			
Temperature	✓		
Glasgow Coma scale (GCS)	✓		✓
Respiratory rate	✓		
Pulse	✓		
Systolic blood pressure	✓		
Diastolic blood pressure	✓		
Oxygen saturation	✓		
Laboratory results			
B-thrombocytes	✓	✓	
B-erythrocytes	✓		
P- partial pressure of oxygen	✓		

P- C-reactive protein (CRP)	✓	✓	✓
B-hemoglobin	✓	✓	
P-albumin	✓	✓	✓
P-creatinine	✓		✓
P-bilirubin	✓		
P-glucose	✓		✓
P-pH	✓		
P-lactate	✓	✓	
B-leucocytes	✓		
B-neutrofilocytes	✓		
B-monocytes	✓	✓	
Procedure			
Urinary catheter	✓	✓	✓

Lastly, we impute the most frequently graded discrete level in our dataset for the methods relying on complete data (NN and LR, respectively). The most frequent grading for each measure is the normal/uncomplicated level, e.g., not smoking, non-alcoholic, a temperature between 36 ad 37.9, a Glasgow Coma Scale (GCS) of 15, a P-glucose between 4 and 7.99, etc. Missing values occur for smoking-, alcohol-, and exercise status, in addition to all vital parameters and laboratory results.

Machine learning models

The dataset contains 138.560 unique hospital admissions (1877 cases of HA-UTI), which is randomly split into 80 % training- and 20 % test sets containing 110.870 (1484 cases of HA-UTI) and 27.690 (393 cases of HA-UTI) unique hospital admissions, respectively.

Seven ML models are trained on the full dataset and our two reduced datasets (Table 1), respectively. As a result, we train, test, and compare 21 ML models (seven ML models for three datasets) for the prediction of patients at risk of HA-UTI. We use LR, naïve BN, and Ds, which support model-specific explainability [37]. We also use a black-box NN in addition to three ensemble methods: RF, AD, and GB. Python's sklearn machine learning package version 1.0.2 (https://scikit-learn.org/) [31] is used to train all our ML models.

The LR models are l2 regularized with the saga solver, with a maximum number of iterations set to 600. The DT models use the categorical and regression tree (CART) [38] and are trained with the Gini impurity criterion (measure for quality of split), best splitter, and a maximum depth of eight. The NN models are a feed-forward architecture, trained using the adam-solver with a max iteration set to 600. Each NN has ten hidden units and three hidden layers in their architecture. The RF models have a maximum of 100 trees (using averaging to improve the decision), are trained with the Gini impurity criterion, and each tree has a maximum depth of eight. The AD models use the Adaboost-SAMME.R algorithm for boosting and a maximum of 100 trees. The GB models uses the Friedman mean-squared-error as the criterion and a maximum of 100 trees.

Model validation

A 10-fold cross-validation is applied over the training set to validate our models and to tune the parameters, whose values are used to control the learning process, based on a mean area under the curve (AUC) over the ten folds.

Test of performance

A receive-operating curve (ROC) is used to analyze the relationship between sensitivity and 1-specificity in the ML models. We also calculate an AUC to compare the performance of our ML models on the test. The test set, unseen for the training phase, is used to test performance. For comparison, we also report on the mean AUC and standard deviation (std) from the 10-fold cross-validation over the training set in soft brackets.

SHapley Additive exPlanation

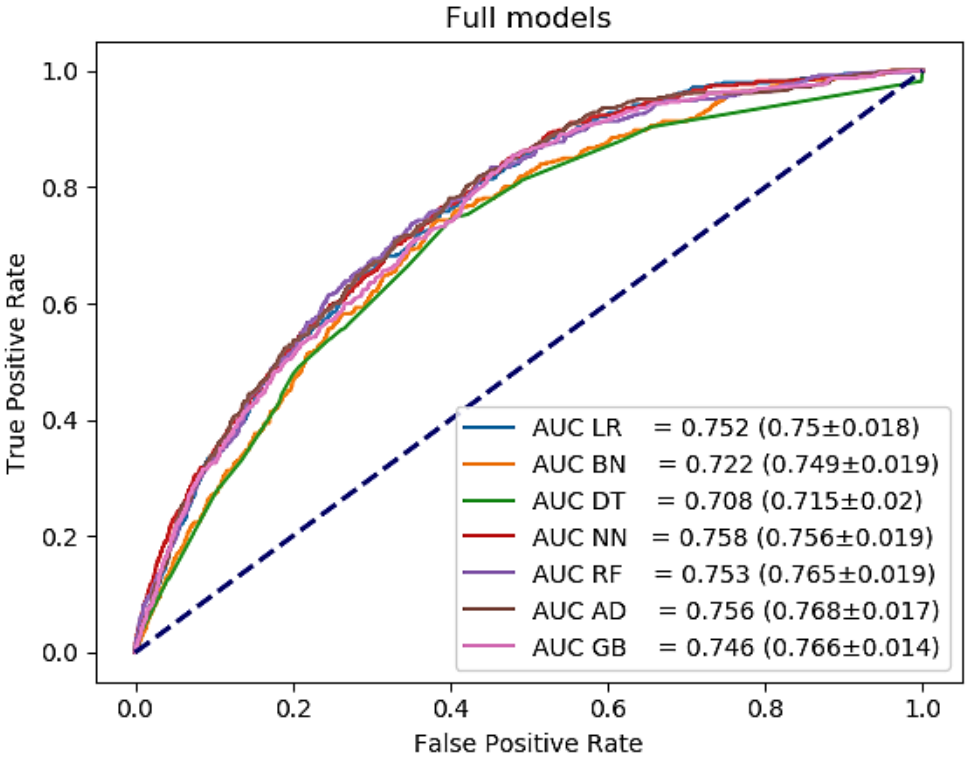
The feature contribution in SHAP is determined by its associated SHAP value, a measure for determining the effect of a feature in a given model prediction [21]. We used a summary plot (*beeswarm*) for population-level model-agnostic explainability and a forceplot for patient-level model-agnostic explainability, available within the SHAP package version 0.41.0 in Python. The KernelExplainer, a weighted linear regression, is used to compute the SHAP values for our models [21]. While the KernalExplainer works without any model-specific assumptions, making it slower than other methods, it is also a generic explainer that works for any ML model type, including naïve BNs, that does not work with other explainers [13]. We include all 393 admissions with patients developing HA-UTI from the test set to produce the summary plot. For the forceplot, we chose randomly one patient with HA-UTI and two without HA-UTI. In population-level model-agnostic explainability, a SHAP summary plot visualizes the extent of each feature's positive- and negative contribution across all predictions, colors their high or low value, and sorts the importance of the features dependent on their contribution. For patient-level explainability, a SHAP forceplot visualizes the extent of each feature's positive- and negative impact on a unique given risk prediction.

Results

Model performance

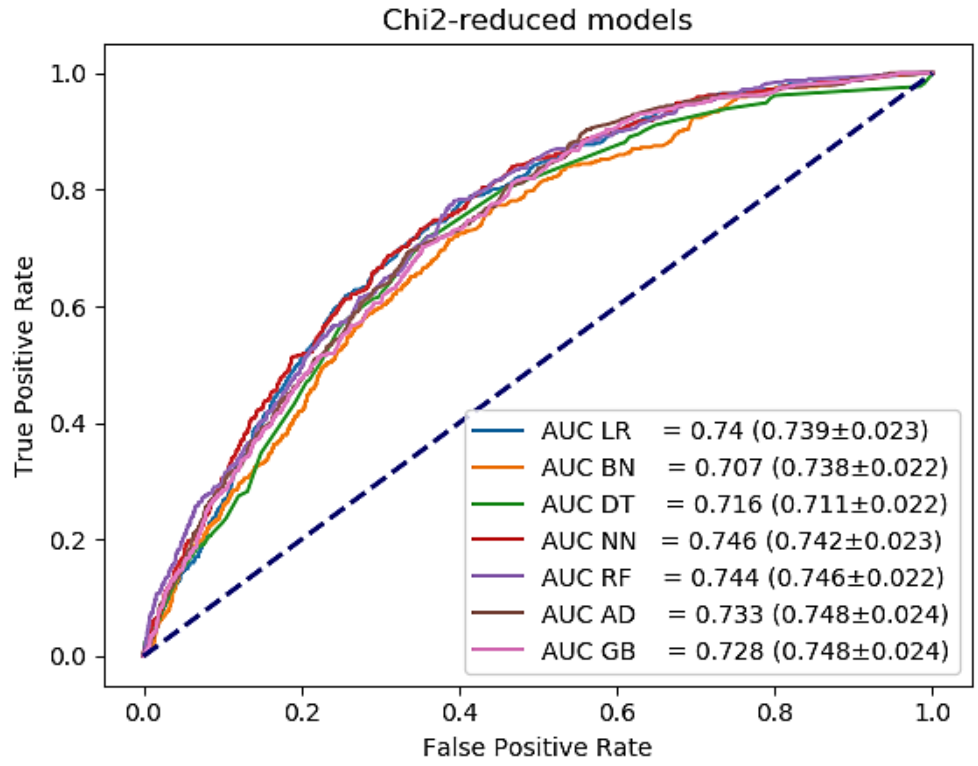
Figure 1 presents the ROC curves of our 21 ML models, the associated AUC from our test, and the mean AUC ±std resulting from the 10-fold cross-validation over the training set.

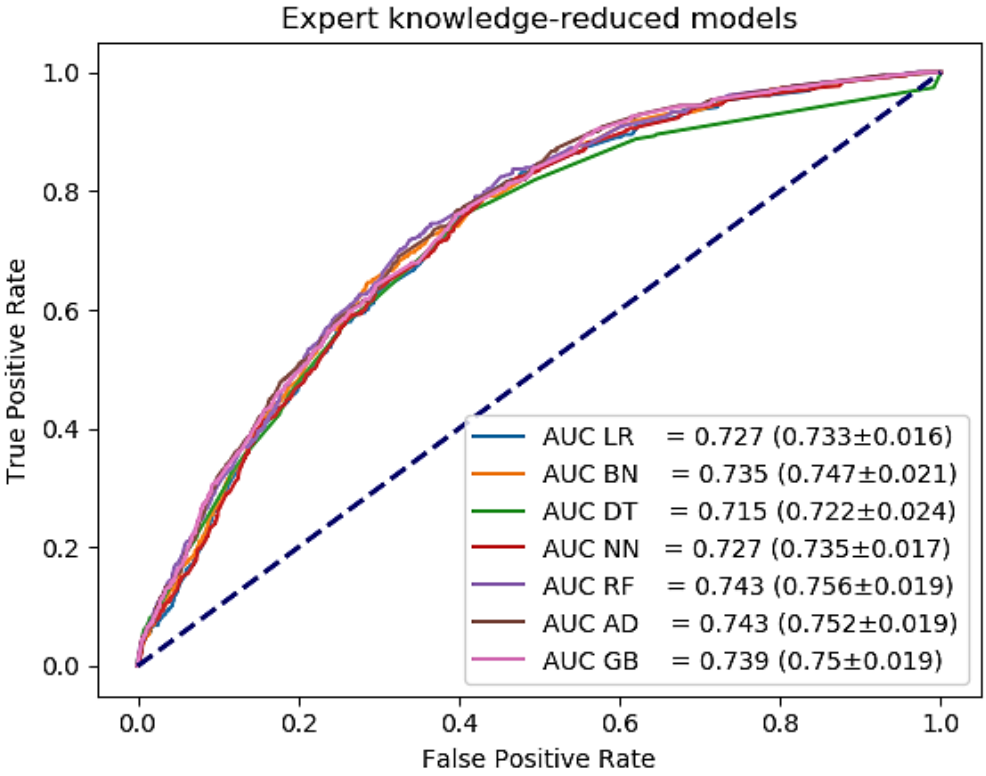
Figure 1: Receiver-operating characteristic curves for the seven ML models over the three different datasets (full, Chi^2 and expert knowledge, respectively)[1]



[17]The receiver-operating characteristic (ROC)-curves for the logistic regression (LR), Bayesian network (BN), and decision tree (DT), Neural Network (NN), Random Forrest (RF), AdaBoost classifier (AD), and Gradient-Boosting

classifier (GB), based on the full-, χ^2-, reduced-, and expert knowledge-reduced-dataset. The LR, NN, RF, AD and GB models have a slight decrease in area under the curve (AUC) between the full dataset and reduced dataset, whereas the DT models seems to reach higher AUC from our feature selection. In addition, the BN reach a higher AUC from the feature selection using expert-knowledge.





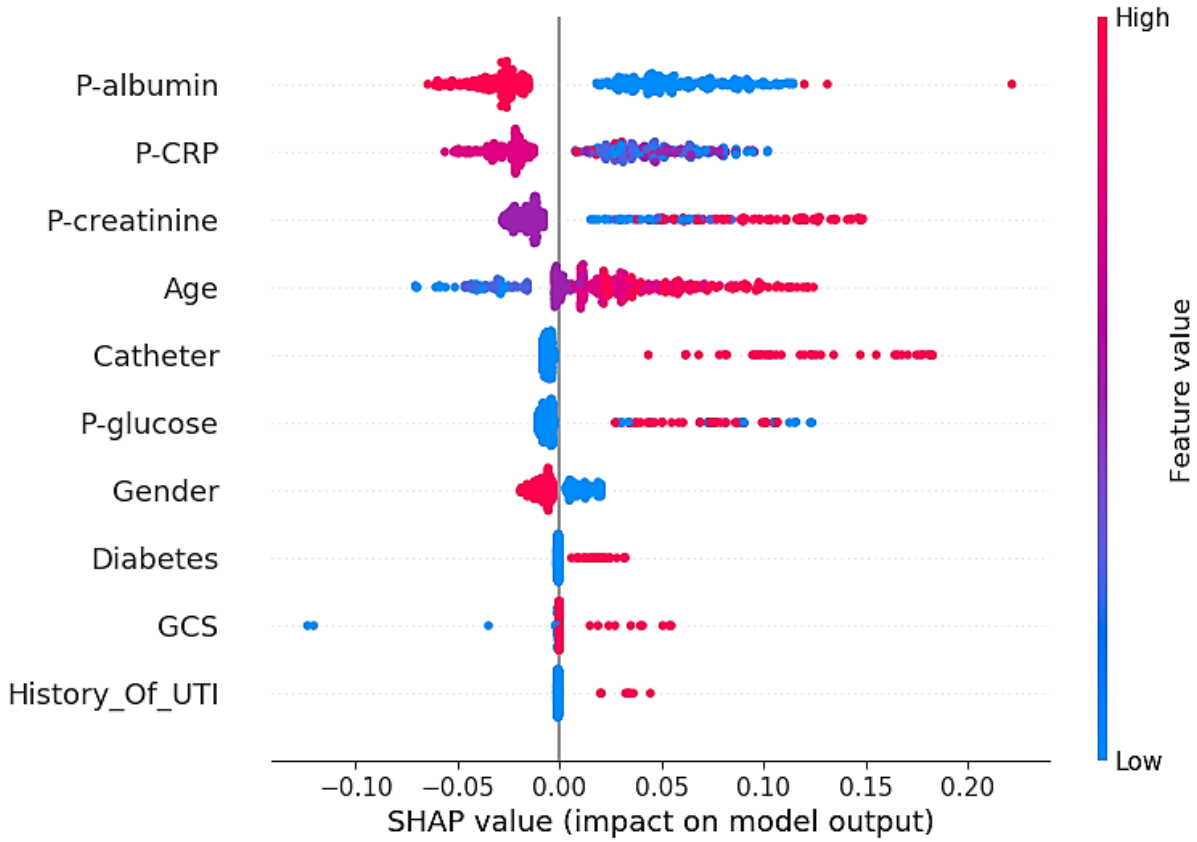
The best-performing model is the NN over the full dataset with the 51 features, reaching an AUC of 0.758 (0.756±0.019), whereto the best-performing model over the reduced datasets is the NN based on the χ^2-feature selection, reaching an

AUC of 0.746 (0.742±0.023). The best-performing ML model that supports model-specific explainability is the LR over the full dataset, reaching an AUC of 0.752 (0.75±0.018), whereto the best-performing model over the reduced datasets is the LR based on the χ^2-feature selection, reaching an AUC of 0.74 (0.739±0.023). However, the BN from the dataset reduced by expert knowledge, which is associated with better knowledge representation between the reduced datasets, has the best performance among all ML models that support model-specific explainability for this dataset, reaching an AUC of 0.735 (0.747±0.021). Therefore, the BN from the dataset reduced by expert knowledge is used to exemplify the SHAP methods for model-agnostic explainability.

Population-level explainability

Figure 2 illustrates the SHAP summary plot for the 393 admissions with patients developing HA-UTI using the BN model trained on the expert knowledge dataset.

Figure 2: SHapley Additive exPlanation summary plots for 393 hospital admissions in which a hospital-acquired urinary tract infection occurred, using the Bayesian Network model based on the expert knowledge-reduced dataset[18]



The contribution of the features is listed from the top down, suggesting that plasma (P) -albumin is the most important feature for predicting patients at risk of HA-UTI, followed by P-creatinine, P-CRP, and age. Urinary catheter has the

[18] The features are ordered from top-down dependent on their contribution. A red dot indicate a high feature value and blue dot indicate a low feature value. The contribution on the prediction is plotted against the x-axis, which is the SHAP value for each feature value from a given hospital admission.

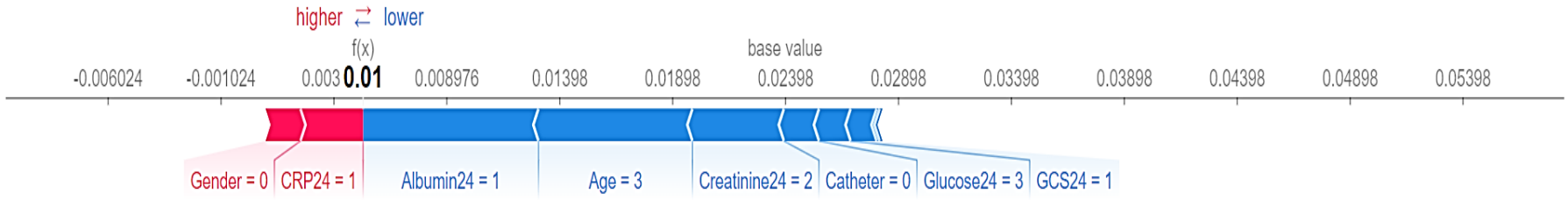
highest contribution on the probability of experiencing HA-UTI while having diabetes, and a history of UTI is the least important features for predicting the risk of HA-UTI, yet with an elevated predicted probability if having these conditions. Moreover, low-graded discrete levels of P-albumin (Supplementary A) are associated with increased probability, whereto high- and low-graded discrete levels of P-CRP and P-creatinine are associated with an increased risk of HA-UTI in the BN model. Moderately increased graded discrete levels of P-CRP and slightly increased graded discrete levels of P-creatinine decrease the probability of HA-UTI. Low age was associated with a lower probability- and higher age was associated with a higher probability of HA-UTI. Female gender slightly increases the risk of HA-UTI, in comparison with the male gender. High-graded discrete levels of GCS (meaning lower scores on the GCS), having diabetes, and a history of UTI may slightly increase the risk of HA-UTI, but these features are often associated with sparse contribution.

Patient-level explainability

Figure 3 presents the SHAP forceplot for the two randomly selected patients without HA-UTI and one randomly selected patient with HA-UTI, using the BN model trained on the expert knowledge dataset.

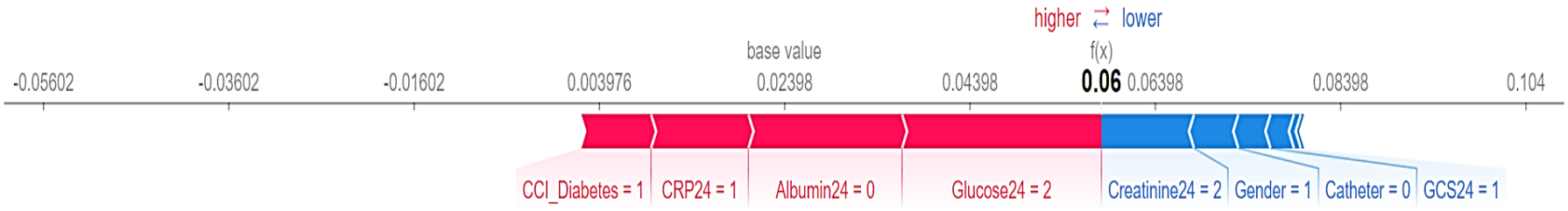
Figure 3: SHapley Additive exPlanation forceplots for three randomly selected individual patients using the Bayesian Network model based on the expert-knowledge reduced dataset

Case #1 without HA-UTI – not identified as a risk patient.[19]



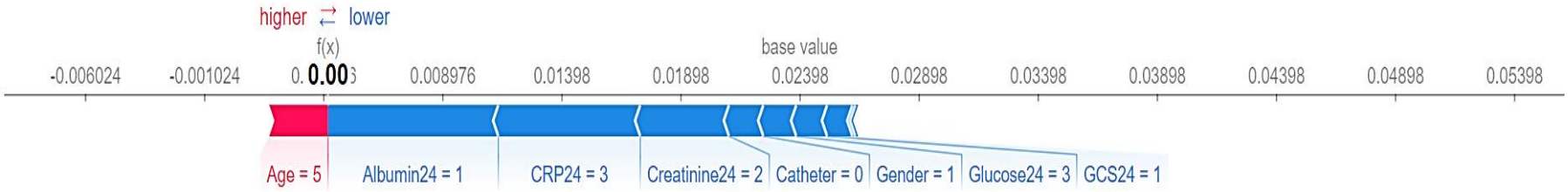
[19] Case #1 is a 50 – 59 years old female. The predicted probability for HA-UTI was lowered at 1.0 % compared with the baseline risk at 2.40%. At admission a high P-glucose at between 20 – 32.99 mmol/L was detected. The P-albumin level was > 48 g/L and P-creatinine level was 105 – 119.99 µmol/L, indicating potential dehydration but otherwise normal kidney function. P-CRP moderately elevated between 8 – 19.99 mg/L. These feature values, in addition to the age and absence of urinary catheter, were the most important contributors to the lowered risk of HA-UTI. The BN model's predicted outcome of HA-UTI was in accordance with the clinical absence of HA-UTI (*true negative*).

Case #2 with HA-UTI – identified as a risk patient.[20]



[20] Case #2 is a 40 – 49 years old male with diabetes. The predicted probability for HA-UTI was elevated at 6.0 % compared with the baseline risk at 2.4%. At admission a moderately high P-glucose between 8 and 11.99 mmol/L together with a normal P-albumin level between 34 and 47.99 g/L. P-CRP was moderately elevated between 8 and 19.99 mg/L. These features caused that the BN model predicted a higher probability for HA-UTI. Male gender, normal Glascow Coma Scale (GCS) score, normal kidney function and absence of a urinary catheter lowered the risk of HA-UTI. The BN model's predicted outcome of HA-UTI was in accordance with the clinical presence of HA-UTI (*true positive*).

Case #3 without HA-UTI – not identified as a risk patient[21]



[21] Case #3 is a 70 – 79 years old male. The predicted probability for HA-UTI was lowered at 0.0 % compared with the baseline risk at 2.4%. At admission a high P-glucose between 20 and 32.99 mmol/L was detected, together with increased P-albumin level > 48g/L and elevated P-CRP between 40 and 199.9 mg/L. Therefore, patient probably started antibiotic treatment that protected against developing an HA-UTI, which together with absence of urinary catheter, may explain the lowered risk of HA-UTI, despite the higher age. The BN model's predicted outcome of HA-UTI was in accordance with the clinical absence of HA-UTI (*true negative*).

The base value in the forceplot is the output if nothing is known for the model, whereto the SHAP values reflect how each feature contributes to reaching the predicted value (f(x)), given the observations. If a feature increases the predicted value f(x), it is colored red in the forceplot. If a feature decreases the predicted value f(x), it is colored blue. The size of each bar in the forceplot results from its SHAP value for the given feature ordered by its contribution, which depends on the graded discrete level (Supplementary A). Only features with a SHAP value above a predefined threshold (by SHAP) are displayed in the SHAP forceplot.

Discussion

In this study, we trained, tested, and compared different ML models for the early identification of patients at risk of HA-UTI, and further supported the explainability of the ML models using the SHAP analysis. We have demonstrated that it is possible to stratify patients within 24h of hospital admission from clinical explainable ML.

The best-performing ML model is based on a NN trained on the full dataset consisting of 51 features. After an automated data-driven- and manual feature selection using expert knowledge, we obtained a high performance of the ML models based on only the ten most meaningful features. Furthermore, we exemplified how the explainability of the BN model was further promoted at both individual patient-level and population-level by the model-agnostic SHAP-method. As a result of this, we provide clinical reasoning of why the ML model reached the given risk prediction.

Meaningfulness of the clinical features

The two different feature selection methods both include age, urinary catheter, P-CRP, and P-albumin as predictors for HA-UTI (Table 1). While age and urinary catheter are well-known risk factors for HA-UTI [32,33], we believe that the agreement between the automated- and manual feature selection on P-CRP and P-albumin reflects the clinical nature of our features, in addition to the ML model design (risk prediction within 24h of admission). For instance, P-CRP [36] and P-albumin [35,36,39] have a broad indication area, being biomarkers for assessing both chronic and acute diseases (such as inflammatory- or malign diseases), which may imply an essential correlation to the patient's more general health status at hospital admission, hence contributing with clinical information in risk assessment in relation to HA-UTI.

For the χ^2 based feature selection, higher χ^2 values are reached for admission cause, the number of comorbidities, and B-lactate. This reflects a higher difference between expected and observed counts, meaning that higher χ^2 values are reached for these features, associated with greater dependence and, as a result, the suggestion of these features as better for predicting HA-UTI. For our expert knowledge-reduced dataset, we also include gender, diabetes, and history of UTI, GCS, and P-glucose, which is not included in the χ^2-reduced datasets (Table 1). We associate gender, diabetes, and history of UTI with well-known risk factors for HA-UTI [32,33,40]. High P-creatinine levels may occur if the patient has impaired kidney function, a well-established risk factor for UTI [35,41]. GCS describes the mental status of the patient, which may be explained by chronic conditions (e.g., mental or behavioral disorders, neoplasms, dementia, or diseases of the nervous system) or acute clinical conditions (e.g., head trauma, cerebral infection, and cardiovascular collapse), which is associated with an elevated risk of urinary tract infections [42]. The P-glucose level may relate to diabetes or being stress-induced, but also as a result of the insulin sensitivity that may be implicated by a fever or as a descriptor for kidney damage, which may consequently increase the risk of experiencing an HA-UTI. Glycemic control for the patient may be critical in preventing HA-UTI [43]. Notably, the χ^2-reduced datasets tend to favor laboratory results, whereto the expert knowledge dataset includes different types of health socio-demographic and clinical data (Table 1).

Balancing performance and clinical explainability

This study adds to the literature's commonly debated trade-off between ML models' performance and explainability [16,17,20,21]. However, we report no significant difference in performance while also pursuing model-specific explainability. This is close in spirit to the suggestions by Rudin [37], who highlights the use of ML models that supports model-specific explainability in the high-stakes healthcare domain and only include the most meaningful clinical features for their purpose. In comparison, the best-performing ML model from Møller et al. [9] in predicting risk within 48h of admission of HA-UTI is their NN model with an AUC of 0.770. However, due to the black-box nature of these models, they recommend the interpretable DT for their prediction of risk, which reached an AUC of 0.709. Our best-performing DT is from the χ^2-reduced dataset reaching an AUC of 0.716. We also consider LR and BN suited for supporting model-specific explainability, which performed better on all datasets compared to the DTs (Figure 1). To further support the explainability of the ML models presented in this study, we demonstrate how to support model-agnostic explainability at the population- and patient-level. A similar approach is found by Deshmukh et al. [20] and Stenwig et al. [13], who

demonstrate the SHAP forceplot and SHAP summary plots for local and global ML model-agnostic explainability in predicting mortality in the ICU. A recently published Danish study by S. Lauritsen et al. [44] also demonstrates the same SHAP methods for the early identification of an acute illness. The SHAP library offers a palette of different visualization techniques and, therefore, different visual explanations for both population- and patient-level model-agnostic approaches. Future studies may explore how different plots contribute differently to the clinical reasoning of the ML models' given result or how the SHAP explanations may differ from other model-agnostic methods, such as the local interpretable model-agnostic explanation (LIME)-methods [45].

Lastly, the SHAP method may be used to inspect and assess the potential clinical implications of the suggestions from an ML model. For instance, when a high P-CRP level for patient 3 (Figure 3) results in a blue bar (negative impact of prediction of risk for HA-UTI) in the SHAP forceplot, it may indicate that some treatment, such as antibiotic therapy, has been given sometime within the hospital admission. Therefore, a high P-CRP level may be a measure for 'given antibiotic therapy' rather than 'higher risk of infection', resulting in the negative impact in the SHAP forceplot. We did not have access to data on antibiotic treatment, which also limits the dataset. We emphasize that future explainable ML models in infection control should still be interpreted with care by the clinician as they may reflect other underlying and clinically significant patterns than perhaps intended.

Study limitations

While the case definition of HA-UTI used in this study is considered remarkably robust [46], the included comorbidities are only based on ICD-8 and ICD-10 codes (Supplementary C), which are sometimes collected for purposes other than documenting disease history. As a result, we may sometimes encounter a negative case of, e.g., diabetes or a history of UTI that should have been detected as a positive, which otherwise could have resulted in better performance or explained a more accurate underlying pattern. For instance, in the summary plot, having diabetes or a history of UTI sometimes increases HA-UTI risk, which is only captured if the comorbidity is included as a positive case in the dataset. If better case definitions for comorbidities become available in the future, the performance test and SHAP methods presented in this study readily apply to such a dataset. Also, as a limitation of our dataset, we impute the most frequently used value for each feature if missingness occurred for the NN and LR because these ML models rely on a complete dataset. However, in an explainable clinical context, this may lead to wrong reasoning for the prediction of risk for HA-UTI, as it may be based on an estimate rather than an actual clinical measure, which may also be reflected in the SHAP analysis. The BN and tree-based ML models do not rely on the imputed dataset.

Perspective

The next step for the clinical explainable ML models presented in this study may be to reach implementation, where clinical experience, confidence, and trust in using ML models for early prediction of patients at risk of an HA-UTI may be gained. In doing so, inspiration may be found in the field of responsible ML in healthcare [12], where the deployment of the ML models is unfolded through various steps with different essential stakeholders, e.g., clinicians, ML developers, lawyers, implementation- and maintenance experts. However, more research is necessary to fully understand the role of ML and explainability in future infection control for HA-UTI to ultimately benefit future patients by avoiding these adverse outcomes.

Conflict of interest

The authors declare no conflict of interest.

Funding source

This study received funding from the Region North Denmark Health Innovation Foundation.

References

[1] Saint S, Meddings JA, Calfee D, et al. Catheter-associated Urinary Tract Infection and the Medicare Rule Changes. 2009;150:877–884.

[2] Saint S. Clinical and economic consequences of nosocomial catheter-related bacteriuria. Am J Infect Control. 2000;28:68–75.

[3] ECDC, EMEA. The bacterial challenge: time to react. 2009.

[4] Conway LJ, Carter EJ, Larson EL. Risk Factors for Nosocomial Bacteremia Secondary to Urinary Catheter-Associated Bacteriuria: A Systematic Review. Urol Nurs. 2015;35:191–203.

[5] Zingg W, Holmes A, Dettenkofer M, et al. Hospital organisation, management, and structure for prevention of health-care-associated infection: A systematic review and expert consensus. Lancet Infect Dis. 2015;15:212–224.

[6] World Health Organisation. WHO guidelines on Hand Hygiene in Health Care. World Heal Organ. 2017;30:64.

[7] Andreessen L, Wilde MH, Herendeen P. Preventing catheter-associated urinary tract infections in acute care: The bundle approach. J Nurs Care Qual. 2012;27:209–217.

[8] Gould IM. Antibiotic policies to control hospital-acquired infection. J Antimicrob Chemother. 2008;61:763–765.

[9] Møller JK, Sørensen M, Hardahl C. Prediction of risk of acquiring urinary tract infection during hospital stay based on machine-learning: A retrospective cohort study. PLoS One. 2021;16:1–16.

[10] Mitchell BG, Ferguson JK, Anderson M, et al. Length of stay and mortality associated with healthcare-associated urinary tract infections: A multi-state model. J Hosp Infect. 2016;93:92–99.

[11] Park JI, Bliss DZ, Chi C-L, et al. Knowledge Discovery With Machine Learning for HospitalAcquired Catheter-Associated Urinary Tract Infections. 2020;38:28–35.

[12] Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019;25.

[13] Stenwig E, Salvi G, Rossi PS, et al. Comparative analysis of explainable machine learning prediction models for hospital mortality. 2022;1–14.

[14] Shaban-Nejad A, Michalowski M, Buckeridge DL, et al. Explainable AI in Healthcare and Medicine. 2020.

[15] Vellido A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Comput Appl. 2020;32:18069–18083.

[16] Stiglic G, Kocbek P, Fijacko N, et al. Interpretability of machine learning-based prediction models in healthcare. Wiley Interdiscip Rev Data Min Knowl Discov. 2020;10:1–13.

[17] Ahmad MA, Teredesai A, Eckert C. Interpretable machine learning in healthcare. Proc - 2018 IEEE Int Conf Healthc Informatics, ICHI 2018. 2018;447.

[18] Burkart N, Huber MF. A survey on the explainability of supervised machine learning. J Artif Intell Res. 2020;70:245–317.

[19] Molnar C, Casalicchio G, Bischl B. Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. Commun Comput Inf Sci. 2020;1323:417–431.

[20] Deshmukh F, Merchant SS. Explainable Machine Learning Model for Predicting GI Bleed Mortality in the Intensive Care Unit. Am J Gastroenterol. 2020;115:1657–1668.

[21] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;2017-Decem:4766–4775.

[22] Pollard TJ, Johnson AEW, Raffa JD, et al. The eICU collaborative research database, a freely available multi-center database for critical care research. Sci Data. 2018;5:1–13.

[23] Condell O, Gubbels S, Nielsen J, et al. Automated surveillance system for hospital-acquired urinary tract infections in Denmark. J Hosp Infect. 2016;93:290–296.

[24] Gubbels S. The development of a national surveillance system for hospital-acquired infections in Denmark The Hospital Acquired Infections Database - HAIBA. 2016.

[25] Sundararajan V, Henderson T, Perry C, et al. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. J Clin Epidemiol. 2004;57:1288–1294.

[26] Hansen BV. Acute admissions to internal medicine departments in Denmark

- studies on admission rate , diagnosis , and prognosis - PhD Dissertation. 2014;

[27] Christensen DH, Veres K, Ording AG, et al. Risk of cancer in patients with thyroid disease and venous thromboembolism. Clin Epidemiol. 2018;10:907–915.

[28] Nielsen PR, Benros ME, Mortensen PB. Hospital contacts with infection and risk of schizophrenia: A population-based cohort study with linkage of danish national registers. Schizophr Bull. 2014;40:1526–1532.

[29] Iguyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157–1182.

[30] Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. London, Edinburgh, Dublin Philos Mag J Sci. 1900;50:157–175.

[31] Pedregosa F, Varoquaux G, Thirion AGVMB. Scikit-learn: Machine Learning in Python. J ofMachine Learn Res. 2011;12:2825–2830.

[32] King C, Garcia Alvarez L, Holmes A, et al. Risk factors for healthcare-associated urinary tract infection and their applications in surveillance using hospital administrative data: A systematic review. J Hosp Infect. 2012;82:219–226.

[33] Redder JD, Leth RA, Møller JK. Analysing risk factors for urinary tract infection based on automated monitoring of hospital-acquired infection. J Hosp Infect. 2016;92:397–400.

[34] Aloush SM, Al Qadire M, Assmairan K, et al. Risk factors for hospital-acquired non-catheter-associated urinary tract infection. J Am Assoc Nurse Pract. 2019;31:747–751.

[35] Zhang Q, Shi B, Wu L. Characteristics and risk factors of urinary tract infection in patients with HBV-related acute-on-chronic liver failure: A retrospective study. Med (United States). 2022;101:E29913.

[36] Yoshida B, Nguyen A, Formanek B, et al. Hypoalbuminemia and Elevated CRP are Risk Factors for Deep Infections and Urinary Tract Infections After Lumbar Spine Surgery in a Large Retrospective Patient Population. Glob Spine J. 2021;13:33–44.

[37] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206–

215.

[38] Breiman L. Classification and regression trees. Breiman L, editor. Boca Raton, Fla: Chapman & Hall; 1984.

[39] Wiedermann CJ. Hypoalbuminemia as surrogate and culprit of infections. Int J Mol Sci. 2021;22.

[40] Storme O, Saucedo JT, Garcia-Mora A, et al. Risk factors and predisposing conditions for urinary tract infection. Ther Adv Urol. 2019;11:19–28.

[41] Scherberich JE, Fünfstück R, Naber KG. Urinary tract infections in patients with renal insufficiency and dialysis - epidemiology, pathogenesis, clinical symptoms, diagnosis and treatment. GMS Infect Dis. 2021;9:Doc07.

[42] Delgado A, Cordero G-G E, Marcos S, et al. Influence of cognitive impairment on mortality, complications and functional outcome after hip fracture: Dementia as a risk factor for sepsis and urinary infection. Injury. 2020;51:S19–S24.

[43] Wang MC, Tseng CC, Wu AB, et al. Bacterial characteristics and glycemic control in diabetic patients with Escherichia coli urinary tract infection. J Microbiol Immunol Infect. 2013;46:24–29.

[44] Lauritsen SM, Kristensen M, Olsen MV, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat Commun. 2020;11:1–11.

[45] Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. NAACL-HLT 2016 - 2016 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol Proc Demonstr Sess. 2016;97–101.

[46] Sips ME, Bonten MJM, Van Mourik MSM. Automated surveillance of healthcare-associated infections: State of the art. Curr Opin Infect Dis. 2017;30:425–431.

STUDY III: A MACHINE LEARNING-GUIDED STRATEGY FOR REPLACEMENT OF PERIPHERAL VENOUS CATHETERS IN THE PREVENTION OF HOSPITAL-ACQUIRED BACTEREMIA: A COMPARATIVE COST-EFFECTIVE ANALYSIS

Rune Sejer Jakobsen[a, b], Thomas Dyhre Nielsen[c], Jakob Juul Christensen[d, e], Sabrina Storgaard Sørensen[d,e], Peter Leutscher [a, e] and Kristoffer Koch[a, f]

a) Centre for Clinical Research, North Denmark Regional Hospital, Hjørring, Denmark
b) Business Intelligence and Analysis, North Denmark Region, Denmark
c) Department of Computer Science, Aalborg University, Denmark
d) Danish Center for Health Services Research, Aalborg University, Denmark
e) Department of Clinical Medicine, Aalborg University, Denmark
f) Department of Clinical Microbiology, Aalborg University Hospital, Denmark

The paper is submitted and has been made available for the Ph.D. committee.

Not available in the published version of the dissertation.

APPENDIX A

THE METHODOLOGICAL CONSIDERATIONS FOR CONSTRUCTING THE DECISION TREE

An early cost-effectiveness analysis compares the ML model's cost-effectiveness with standard care. The decision tree portion containing the ML model follows a test-based approach by Tamlyn et al. [121]. First, the ML model divides into two paths, reflecting the model's probability of predicting high HAB risk. The tree further splits into TP and FP cases for high-risk predictions, acknowledging prediction errors. It then branches into scenarios of PVC initiation during admission. If PVC is initiated, the tree explores the likelihood of preventive treatment by the physician, reflecting the reality that not all patients will receive such care. If preventive treatment is provided, the tree splits again to reflect the probability of HAB from PVC use and whether HAB develops despite the treatment. This process repeats for FP cases. The tree is simplified for usual care to represent the probability of PVC initiation and subsequent HAB development during hospital admission.

PROBABILITIES

The performance test probabilities for the ML model were determined based on the best-performing model's TP, TN, FP, and FN rates. The probability of initiating preventive treatment was set at 99% for both TP and FP, anticipating that physicians would typically follow ML recommendations. Approximately 14% of all HAB cases can be linked to the use of PVC [8,21], and around 70% of all hospital admissions involve a PVC [119]. The probability of contracting HAB under usual care is equivalent to that under the ML model.

COSTS

The ML model is expected to be relevant to approximately 69,280 patients per year, half the study cohort over a two-year period. From a Danish hospital sector perspective, the analysis considers costs related to ML model development, implementation, maintenance, staff training, preventative treatment, frequent PVC changes, PVC replacements during admission, HAB treatment, and general admission costs. Initial and capital costs were spread across the total annual patient count, with the ML model expected to be in service for five years.

Preventative treatment costs factored in physician and nurse salaries, their time spent evaluating high-risk patients and changing PVC, the price of catheters, and catheter replacement frequency. The monthly effective working hours for nurses and

physicians were assumed to be 94, consistent with recommendations from The Danish National Health Technology Council [124]. Based on estimates from the Business Intelligence and Analysis Unit in North Denmark Region, a cumulative cost estimate was used for HAB occurrences during hospital stays.

EFFECTIVENESS

The effectiveness measure was determined by integrating the probability of death from a HAB infection into the terminal nodes of the decision tree, where HAB development was indicated. This approach facilitated the computation of the expected effectiveness, representing the average likelihood of dying from a HAB-related infection under the ML model and usual care alternatives. In the study cohort, the probability of mortality due to a HAB infection was 26.89%. This effect measure was inversed in the ICER, considering death from HAB as unfavorable.

ISSN (online): 2246-1302
ISBN (online): 978-87-7573-645-4

AALBORG UNIVERSITY PRESS