Dissertation

Performance and User Acceptance of a Machine Learning-Based Delirium Risk Stratification Tool in Clinical Routine

submitted by

Stefanie JAUK, MSc.

for the Academic Degree of
Doctor of Philosophy
(PhD)

at the
Medical University of Graz
Institute for Medical Informatics, Statistics and Documentation

under the supervision of
Univ.-Prof. Dr. med. Stefan SCHULZ

2021

DISSERTATION COMMITTEE

Univ.-Prof. Dr. med. Stefan Schulz
Institute for Medical Informatics, Statistics and Documentation
Medical University of Graz

Univ.-Prof. Dipl.-Ing. Dr. Andrea Berghold
Institute for Medical Informatics, Statistics and Documentation
Medical University of Graz

Univ.-Ass. Priv. Doz. Mag. Dr. Alexander Avian
Institute for Medical Informatics, Statistics and Documentation
Medical University of Graz

Mag. Dr. Diether Kramer
Steiermärkische Krankenanstaltengesellschaft m.b.H. (KAGes), Graz

DECLARATION

I hereby declare that this thesis is my own original work and that I have fully acknowledged by name all of those individuals and organisations that have contributed to the research for this thesis. Due acknowledgement has been made to all other material used. Throughout this thesis and in all related publications I followed the "Standards of Good Scientific Practice and Ombuds Committee at the Medical University of Graz".

Stefanie Jauk
June, 2021

DISCLOSURES

Part of this thesis is based on the following publications:

- Jauk, S., Kramer, D., Großauer, B., Rienmüller, S., Avian, A., Berghold, A., Leodolter, W. and Schulz, S. (2020). Risk Prediction of Delirium in Hospitalized Patients Using Machine Learning: An Implementation and Prospective Evaluation Study. *Journal of the American Medical Informatics Association*, 27(9), pp.1383–1392.

  According to its publication rights, the Oxford University Press allows to include the article in full or in part in a thesis or dissertation, provided that this is not published commercially.[1]

- Jauk, S., Kramer, D., Avian, A., Berghold, A., Leodolter, W. and Schulz, S. (2021). Technology Acceptance of a Machine Learning Algorithm Predicting Delirium in a Clinical Setting: a Mixed-Methods Study. *Journal of Medical Systems*, 45(48).

  This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as appropriate credits are given to the original author(s) and the source, a link to the Creative Commons licence is provided, and any changes are indicated.[2]

Diether Kramer, Werner Leodolter, Susanne Rienmüller and Birgit Großauer are affiliated with Steiermärkische Krankenanstaltengesellschaft m.b.H. (KAGes); Stefan Schulz, Andrea Berghold and Alexander Avian are affiliated with the Institute of Medical Informatics, Statistics and Documentation of the Medical University of Graz. All co-authors have explicitly agreed to the use of the published data in this thesis.

In addition, the following contributions were made:

- The machine learning models under evaluation in this thesis were developed by Diether Kramer. Günther Schreier, Dieter Hayn and Sai Veeranki from the Austrian Institute of Technology, Franz Quehenberger from the Medical University of Graz as well as I provided support in the development as part of a research project carried out by CBmed GmbH.

1 https://academic.oup.com/journals/pages/access_purchase/rights_and_permissions/publication_rights

2 https://creativecommons.org/licenses/by/4.0/

- The web application of the delirium risk stratification tool was developed by Diether Kramer using the R package *shiny* (Chang et al., 2020), and it was described by Veeranki et al. (2018).
- The software design, the integration of the machine learning models into the risk stratification algorithm and the implementation in the hospital information system were mainly carried out by Diether Kramer, supported by the team of Medical Informatics and Process Management (MIP) from KAGes.
- Activities of implementation and the training sessions for healthcare professionals in the participating hospitals were conducted by Diether Kramer and me.
- Clinical advice for the development of the tool and for the implementation process was provided by Günther Stark, Christian Jagsch, Herbert Wurzer, Hubert Hauser, Ewald Tax, Susanne Rienmüller, Birgit Großauer and many other clinicians affiliated with KAGes.

PUBLICATIONS

In addition to the two mentioned publications, I published the following articles as first author during the PhD programme (in chronological order):

- Jauk, S., Kramer, D. and Leodolter, W. (2018). Cleansing and Imputation of Body Mass Index Data and Its Impact on a Machine Learning Based Prediction Model. *Studies in Health Technology and Informatics*, 248, pp.116–123.
- Jauk, S., Kramer, D., Schulz, S. and Leodolter, W. (2018). Evaluating the Impact of Incorrect Diabetes Coding on the Performance of Multivariable Prediction Models. *Studies in Health Technology and Informatics*, 251, pp.249–252.
- Jauk, S., Kramer, D., Quehenberger, F., Veeranki Sai Pavan, K., Hayn, D., Schreier, G. and Leodolter, W. (2019). Information Adapted Machine Learning Models for Prediction in Clinical Workflow. *Studies in Health Technology and Informatics*, 260, pp.65–72.
- Jauk, S., Kramer, D., Stark, G., Hasiba, K., Leodolter, W., Schulz, S. and Kainz, J. (2019). Development of a Machine Learning Model Predicting an ICU Admission for Patients with Elective Surgery and Its Prospective Validation in Clinical Practice. *Studies in Health Technology and Informatics*, 264, pp.173–177.
- Jauk, S. (2021). Reply to Rousseau and Tierney. *Journal of the American Medical Informatics Association*, 28(3), pp.666–667.

ACKNOWLEDGEMENTS

Finishing a PhD is not a one-woman show; I have received a great deal of support and assistance while working on this thesis over the last years.

Without any doubts, Diether Kramer deserves major recognition for supporting me in this process. Thank you, Diether, for providing your experience and expertise, and for offering an exciting but also pleasant working environment. Without you, this PhD project would not have been possible.

Special thanks go to my supervisor, Stefan Schulz, who gave me the opportunity to start my research in this field. Thank you, Stefan, for all our enriching discussions and especially for your ongoing improvement of my scientific texts.

I also want to thank Andrea Berghold and Alexander Avian from the Institute of Medical Informatics and Statistics who supervised my work and provided their expertise. The entire institute has always been an open place for me to discuss, raise questions and receive advice.

Regarding the organisational support, I want to thank the PhD Program Advanced Medical Biomarker Research (AMBRA) of the Medical University of Graz and Barbara Obermayer-Pietsch for guiding us PhD students with her kindness. In addition, my thanks go to the CBmed GmbH and KAGes m.b.H. for providing financial support for my work.

Most importantly, I would like to acknowledge all my friends and my family who have been supporting me over the last years. In particular, I would like to thank my partner Richard, who always provided a look from outside when reading my texts, and who always critically questioned my statements. Thank you, Richard, for supporting my decision to start this PhD and for always believing in me finishing it.

Being the first one in our family to reach this level of education is not self-evident, and thus the biggest thanks go to my parents who supported every step of my formal and informal education and who never stopped believing in me.

CONTENTS

ABBREVIATIONS

Adaboost	Adaptive boosting
ATC	Anatomical therapeutic chemical
AUROC	Area under the receiver operating characteristic
AutoML	Automated machine learning
EHR	Electronic health record
CAM	Confusion assessment method
CART	Classification and regression trees
CI	Confidence interval
GAN	Generative adversarial network
GBM	Gradient boosting machine
GLM	Generalized linear model
HIS	Hospital information system
ICD	International classification of diseases
ICU	Intensive care unit
LIME	Local interpretable model-agnostic explanations
LKH	Landeskrankenhaus (*Federal state hospital*)
LOINC	Logical observation identifiers names and codes
LOS	Length of stay
LR	Logistic regression
LSTM	Long short-term memory
NLP	Natural language processing
RF	Random forest
ROC	Receiver operating characteristic
SVM	Support vector machine
TAM	Technology acceptance model
TAM2	Technology acceptance model 2

LIST OF FIGURES

LIST OF TABLES

ZUSAMMENFASSUNG

Im klinischen Alltag ist es unerlässlich, Patient*innen mit einem erhöhten Risiko für lebensbedrohliche Krankheiten zu identifizieren, um so früh wie möglich präventive Maßnahmen treffen können. Der Einsatz von klinischen Prognosemodellen ermöglicht es Patient*innen in Risikogruppen einzuteilen, wodurch das Gesundheitspersonal in klinischen Entscheidungsprozessen unterstützt werden kann.

Aufgrund zunehmender Datenmengen in elektronischen Gesundheitsakten wurden über die letzten Jahre vermehrt Prognosemodelle mithilfe von Methoden des maschinellen Lernens entwickelt. Der wesentliche Vorteil dieser Methoden besteht darin, dass keine zusätzlichen Daten erfasst werden müssen, und damit ein flächendeckender Einsatz ohne den Aufwand zusätzlicher personeller Ressourcen im klinischen Alltag ermöglicht wird.

Trotz der ausgezeichneten Ergebnisse dieser Modelle auf Testdatensätzen wurden bisher nur wenige Modelle des maschinellen Lernens in den klinischen Einsatz gebracht. Aus diesem Grund sind sowohl die klinische Prognosegenauigkeit als auch die Akzeptanz durch das klinische Personal weitgehend unerforscht.

Das Ziel dieser Dissertation war die Evaluierung einer Anwendung basierend auf den Methoden des maschinellen Lernens im klinischen Alltag. Der Zweck der Anwendung ist die Vorhersage eines Delirs, ein akuter Verwirrtheitszustand, welcher oft mit hohen Komorbiditäten und Sterblichkeitsraten einhergeht. Die Evaluierung umfasste die folgenden drei Aspekte: (1) die Prognosegenauigkeit während einer siebenmonatigen Pilotstudie; (2) die Akzeptanz der Anwendung durch das klinische Personal; und (3) die Langzeit-Prognosegenauigkeit in fünf Krankenhäusern in der Steiermark (Österreich).

Die Ergebnisse der Evaluierung zeigten sowohl in der Pilotstudie als auch in der Langzeituntersuchung eine stabile Prognosegenauigkeit im klinischen Alltag für internistische und chirurgische Patient*innen. Die Risikovorhersagen des Algorithmus waren in hoher Übereinstimmung mit den Einschätzungen der klinischen Expert*innen für eine Stichprobe von Patient*innen der Allgemeinen Inneren Medizin und Gastroenterologie.

Das klinische Personal bewertete die Nützlichkeit, Handhabbarkeit und Ergebnisqualität der Anwendung positiv, und begrüßte insbesondere die schnelle und automatisierte Risikovorhersage. Die Häufigkeit der Nutzung wurde jedoch vom klinischen Personal als eher gering eingestuft. Zukünftige Implementierungsprozesse sollten daher Informationsveranstaltungen und Bekanntmachungen intensiver forcieren.

Im Rahmen der Evaluierung zeigten sich auch Schwächen der auf Methoden des machinellen Lernen basierenden Anwendung. Für Patient*innen der Abteilung für Kardiologie konnte ein Delir nur mit einer vergleichsweise geringen Präzision vorhergesagt werden. Zudem war die Evaluierung durch eine unvollständige Delir-Dokumentation im Krankenhausinformationssystem limitiert. Diverse Ansätze werden diskutiert, um den Einfluss dieser Limitierung für zukünftige Analysen zu reduzieren.

Diese Dissertation gibt neue Einblicke in den Einsatz von maschinellem Lernen im klinischen Alltag und zeigt diverse Stärken und Schwächen einer solchen auf. Die Ergebnisse veranschaulichen sowohl die hohe prädiktive Genauigkeit für die Vorhersage eines Delirs als auch die positive Akzeptanz der Anwendung durch das klinische Personal. Obwohl sich diese Arbeit gezielt auf die Risikovorhersage eines Delirs richtet, können die Erkenntnisse zukünftig auch für die Evaluierung und kritische Beurteilung von Modellen des maschinellen Lernens in anderen klinischen Anwendungsfeldern herangezogen werden.

ABSTRACT

In clinical routine, early identification of patients with life-threatening risks is crucial in order to initiate preventive actions as quickly as possible. Clinical prediction models stratify patients according to their risk and thus support healthcare professionals in their decision-making. Owing to the increasing amount of clinical data stored in electronic health record (EHR) systems, numerous machine learning-based prediction models have been developed over the last years. A main advantage of combining machine learning and EHR data is that no additional information needs to be assessed, which saves resources and allows for routine risk stratification in hospitals.

Despite demonstrations of outstanding prognostic performance in test data sets, only few machine learning models have been implemented in clinical settings. Therefore, little is known about their clinical performance and their acceptance by clinicians.

The goal of this thesis was to evaluate a machine learning-based risk stratification tool in clinical routine. The predicted outcome of the tool is delirium, a syndrome of acute confusional state with high morbidity and mortality in hospitalised patients. The evaluation addressed three aspects: (1) the prospective performance of the delirium risk stratification algorithm in a seven-months pilot study; (2) the technology acceptance of the tool by healthcare professionals; and (3) the long-term performance when implemented in five hospitals across the Austrian region of Styria.

The results demonstrate that the algorithm achieved a stable performance for internal medicine and surgical patients in clinical routine during a pilot study and in the long term. Delirium risk predictions by the algorithm were in high agreement with risk ratings by clinical experts for a sample of general internal medicine and gastroenterology patients. Overall, healthcare professionals rated the usefulness, ease of use and output quality positively and appreciated the automatic and fast prediction. However, the reported use of the tool was still low and more promotion and training sessions will be needed in future deployments.

The evaluation also revealed weaknesses of the machine learning-based tool, e.g. a decrease in performance when applied to a cardiology department with a more complex patient cohort. In addition, a low observed incidence of delirium in the EHR data limited the evaluation, but ways to overcome this limitation in future are discussed.

To conclude, this thesis provides new insights into the clinical performance of a machine learning-based risk stratification tool and illustrates its strengths and

weaknesses. It demonstrates the high predictive performance of machine learning-based delirium prediction and the positive acceptance by healthcare professionals. Even though the focus of this thesis was the prediction of delirium, the results will support the evaluation and critical appraisal of machine learning models for different clinical outcomes in future.

1

INTRODUCTION

Artificial intelligence and particularly machine learning have been a constant in medical informatics research over decades (Coiera, 1996; Peek et al., 2015). The prediction of clinical outcomes, one of many applications of machine learning in medicine, has gained much attention since well-known companies have started developing machine learning models (Rajkomar, Dean, & Kohane, 2019). This has given rise to various clinical prediction models for a broad range of clinical use cases, which have achieved high predictive performance in retrospective data sets.

The goal of clinical prediction models is to support healthcare professionals in risk prediction and decision-making, and thus improve patient care. In the past few years, several machine learning-based prediction models have been published which were trained on data from electronic health record (EHR) systems (Goldstein et al., 2017). The advantage of using EHR data as an input for prediction modelling is the big amount of already available longitudinal clinical data.

For EHR-based risk prediction no additional data need to be assessed, which saves resources in hospitals and allows routine hospital-wide risk stratification. The risk prediction is free of effort for healthcare professionals, which further leads to a higher technology acceptance (Davis, Bagozzi, & Warshaw, 1989).

Although various machine learning-based models have been developed, few of them have ever been deployed to support healthcare professionals in clinical routine (He et al., 2019; Vollmer et al., 2020). As a consequence, only few studies have addressed the model performance in dynamic decision-making situations, implementation processes and the acceptance of the models by healthcare professionals (Islam et al., 2018).

In contrast to rule-based prediction models, machine learning models are able to identify more complex patterns in data, which makes their interpretation more complicated. When implemented in clinical settings, unexpected behaviour of the models may occur and the performance may differ from the results when applied to retrospective data (Amarasingham et al., 2014). Therefore, an ongoing evaluation of these models in clinical settings is crucial to ensure a high performance of prediction models and patient safety (Magrabi et al., 2019).

For a successful clinical deployment, the predicted outcome of the models needs to be actionable and controllable (Bates et al., 2014). A promising use case for machine

learning-based prediction modelling is the occurrence of delirium in hospitalised patients.

Delirium is a syndrome of acute confusional state with an acute decline in attention and cognitive functioning (Inouye, Westendorp, & Saczynski, 2014). The occurrence of delirium is associated with higher mortality and morbidity, but it is preventable in many cases using non-pharmacological interventions. It is therefore highly rewarding to identify patients at high risk of delirium as early as possible. However, as the aetiology of delirium is multifactorial (Inouye, 2006) it remains an open challenge to accurately identify high-risk patients in clinical routine.

1.1 MOTIVATION AND AIM

The aim of this thesis was to evaluate a machine learning-based risk stratification tool in a clinical setting. Since 2016, machine learning models predicting the risk of delirium had been developed, and were finally integrated into a risk stratification algorithm. Starting in 2018, the risk stratification algorithm has been deployed as a decision support tool in the hospital information system (HIS) of a public hospital network in Austria.

This thesis addressed a comprehensive evaluation of the delirium risk stratification tool in clinical routine. The work was divided into three parts addressing the following aims:

1. The first aim of this thesis was to prospectively evaluate the performance of the delirium risk stratification algorithm in a clinical setting. In a seven-month pilot study, the performance of the algorithm was assessed using state-of-the art measures for discrimination and calibration. Besides, risk predictions of the algorithm were compared with the risk ratings of experts to clinically validate the results.
2. The second aim was to gain knowledge on how healthcare professionals perceive machine learning-based risk stratification tools. The technology acceptance by users was evaluated in a mixed methods study including expert group meetings and questionnaire assessments.
3. The third aim of this thesis was to evaluate the long-term performance of the risk stratification algorithm in a multicentre setting. After the pilot study, the tool was deployed in several hospitals of the hospital network in 2019. Prospective predictions of the algorithm were analysed for five hospitals treating different patient cohorts in order to determine the stability of the performance.

1.2 STRUCTURE OF THE THESIS

Chapter 1 provides an introduction to clinical prediction models using machine learning and illustrates the aim of this thesis.

Chapter 2 presents the scientific background of the main aspects of the work. This includes an introduction to machine learning-based prediction models (Section 2.1), barriers to the deployment of machine learning models (Section 2.2), and their evaluation in healthcare (Section 2.3). Furthermore, clinical aspects of delirium are discussed (Section 2.4), and previously developed prediction models for delirium are reviewed (Section 2.5).

Chapter 3 summarizes the main methods used for evaluation, including an overview of performance measures (Section 3.1), a description of the expert group (Section 3.2) and the identification of delirium patients in EHR data (Section 3.3). Finally, the chapter demonstrates the overall study design (Section 3.4).

Chapter 4 describes the main material of evaluation, i.e. the delirium risk stratification tool. This chapter is divided into (i) the development of the machine learning models (Section 4.1), (ii) their integration into the risk stratification algorithm (Section 4.2), and (iii) the presentation of the predicted delirium risk to healthcare professionals (Section 4.3).

The following chapters, Chapter 5, Chapter 6 and Chapter 7, address the three aims of this thesis. Each chapter presents an in-depth introduction, specific methods addressing the respective research question, results and limitations.

General aspects of the thesis are discussed in Chapter 8. The chapter addresses the main findings of the evaluation (Section 8.1) and problems encountered when identifying delirium patients in EHR data (Section 8.2). It provides examples of how to overcome barriers of deployment, and discusses changes induced in healthcare (Section 8.3). Moreover, the chapter highlights strengths and weaknesses of machine learning models (Section 8.4) and the clinical benefit of such tools (Section 8.5). Finally, it provides an outlook for future work to be explored (Section 8.6).

The last chapter, Chapter 9, presents a conclusion of this thesis and final remarks.

2

BACKGROUND

2.1 RISK PREDICTION IN HEALTHCARE

For decision-making in everyday clinical practice, healthcare professionals need to make various diagnostic and prognostic predictions for patients. While diagnostic prediction estimates the probabilities of a disease being present, prognostic prediction refers to the prediction of possible outcomes of a disease and the frequencies of events such as death, survival, cure or complication (Laupacis et al., 1994).

Both types of predictions are essential to make decisions concerning further screening or preventive and therapeutic interventions. In inpatient settings, resources for preventive actions are often limited. Thus, targeting patients at high risk can help to use existing resources more effectively.

A common way to identify patients who are likely to benefit from preventive actions is the use of clinical prediction models (Steyerberg, 2019). Prediction models aim to inform clinicians about patients' diagnostic or prognostic risks. Publications on clinical prediction models have increased over the past years, with most of them using evidence-based data for prediction.

Clinical prediction models have usually been developed out of large cohort studies or based on established clinical guidelines (Goldstein et al., 2017); risk predictions have been made with few predictors which have causal impacts on the predicted outcomes. One of the most famous prediction models in healthcare is the Framingham risk score, which estimates the ten-year cardiovascular risk (D'Agostino et al., 2008). This score has been widely accepted for quantifying the overall cardiovascular risk for patients and for identifying risk factors.

Although numerous risk scores have been recommended by medical associations and integrated into guidelines, many of them are not routinely used. Barriers for their adoption include uncertainty of the target population or the need for additional interventions (Müller-Riemenschneider, 2010). They often lack accuracy, and their population-based risk estimates often differ from individual risks of patients.

At the beginning of this century, Leo Breiman illustrated the need of alternative approaches in addition to the exclusive approach of using stochastic data models based on theory and assumptions (Breiman, 2001b). According to him, algorithmic approaches such as neural networks and decision trees present an alternative and

could help solving complex prediction problems, e.g. when aggregating over a large set of models instead of using a single model.

2.1.1 *Machine Learning Methods for Clinical Prediction Models*

Healthcare is a multifaceted field due to the complexity of diseases, the heterogeneity of outcomes and the variety of diagnostic and therapeutic procedures (Lee & Yoon, 2017). Massive amounts of data are routinely collected in EHR systems. They are constituted by clinical narratives, medical images, diagnostic data increasingly including omics data, and other structured and unstructured data. As human cognition puts limits to the analysis of big amounts of data, the use of algorithms presents one way to overcome these limitations (Topol, 2019).

Machine learning, also referred to as statistical learning (Hastie, Tibshirani, & Friedman, 2009), uses many different data points to predict outcomes with classification and regression techniques. The goal is to learn complex patterns from training data and develop models that generalise well to unseen test data.

There are many definitions of machine learning, and there is no clear boundary between machine learning models and non-machine learning models. Recent discussions suggest to rather classify prediction models based on their ability to model linearity or non-linearity, or to rank them by complexity (Bian et al., 2019). Machine learning used for clinical prediction models can be thought of conceptually as a modelling framework tailored to large or complicated predictor sets.

With the growing amount of routinely collected EHR data, data for modelling can be retrieved directly from clinical information systems. Their availability combined with machine learning methods facilitates the development of prediction models and further personalises patient care (Parikh, Kakad, & Bates, 2016). The machine learning approach does not require additional data entry or calculation, which is usually the case for risk prediction with rule-based scores.

Various attempts have been made with machine learning to predict outcomes in oncological and cardiovascular diseases (Islam et al., 2018), as well as in neurological diseases (Jiang et al., 2017). Such clinical prediction models have been published with promising results, e.g. predicting 30-day hospital readmission (Hao et al., 2015), the first cardiovascular event over ten years (Weng et al., 2017) or short-term mortality among patients with chemotherapy (Elfiky et al., 2018).

Depending on the underlying data and the predicted clinical outcome, different machine learning methods are used for modelling. Clinical prediction problems can be modelled using supervised learning methods, if annotated data are available, or

by unsupervised learning methods, if data are unlabelled (Weng, 2020). Supervised learning can be further divided into classification methods for the prediction of outcome classes, or regression methods, when predicting continuous outcomes.

No single machine learning method performs best for all possible data sets. The recommended methods for supervised class prediction include, among others, generalized linear models (GLM), support vector machines (SVM) and tree-based methods (Weng, 2020).

GLM, in particular logistic regression for binary outcomes, is widely used for clinical prediction models (Reddy & Li, 2015; Steyerberg, 2019). GLM is able to incorporate categorical and continuous predictors, non-linear transformations and interaction terms. Linear methods are often used in combination with regularization techniques in order to improve the generalisability and prevent overfitting on the training data, for instance lasso or ridge regression (Hastie, Tibshirani, & Friedman, 2009). While lasso uses the L1 penalty, which shrinks certain coefficients to zero and removes features, ridge regression uses the L2 penalty, which constricts coefficients but keeps all features.

SVM separates cases of different class labels with the use of hyperplanes in a multidimensional space (Steyerberg, 2019). Both classification and regression problems can be modelled with SVM, and continuous variables can be handled as well as categorical values with numerical representation. Kernel functions are used to introduce non-linearity (Liu, Du, & Feng, 2020). Although SVM often achieves high predictive performance, its interpretability is limited.

Tree-based methods are used for regression and classification problems. They are a good choice for modelling if high-order interactions are expected in big data sets, although information might be partly lost as continuous features have to be categorized (Steyerberg, 2019).

One of the most famous tree-based methods is classification and regression tree (CART), first described by Breiman et al. (1984). CART uses recursive partitioning to construct binary tree models (Steyerberg, 2019). Smaller subgroups of patients are built by finding the statistically optimal split, which results in the maximum separation among two sub-groups. The predictor that causes the largest separation is on top of the tree, and splitting continues until no improvement can be obtained, or subgroups reach a minimum size. Different splitting criteria are available to obtain the optimal model, e.g. Gini impurity index or cross entropy.

2.1.2 *Random Forest Models in Healthcare*

Binary decision trees follow simple presentations and are easy to interpret by humans (Weng, 2020). However, as they may have high variance, techniques such as feature subsampling, boosting or bagging have been used for tree-based methods. Besides adaptive boosting (Adaboost) and gradient boosting machine (GBM), random forest is a commonly applied tree-based method for clinical prediction models.

Random forest, developed by Breiman (2001a), is an ensemble method building a big collection of de-correlated decision trees and averaging them (Hastie, Tibshirani, & Friedman, 2009). It is based on bagging, a technique to reduce the variance of an estimated prediction function. Random forest reduces the correlation between trees, as it uses a random selection of input variables when growing the trees. It further improves the variance reduction of bagging, and reduces overfitting of single decision trees. Random forest models are able to account for non-linear relationships in the data.

The use of random forest for clinical prediction models has increased over the past years, and there are various publications demonstrating good predictive performance in healthcare (Bihorac et al., 2019; VanHouten et al., 2014; Weng et al., 2017).

2.1.3 *Predictive Modelling Using Electronic Health Records*

The secondary use of EHR data has grown in the past years, but no consensus has been found whether the reuse of clinical data satisfies research standards. While some stand in for using data only for the collected purpose, others argue that data quality is sufficient as soon as specific goals can be reached (Weiskopf & Weng, 2013).

The quality of EHR data ranges widely between different EHR systems (Weiskopf & Weng, 2013). EHR data can be incomplete if data are available only on paper records or collected in other institutions outside the hospital network. Values can be incorrect, e.g. due to wrong data entry, coding errors because of uncertainty in terminologies (Stausberg et al., 2008) or upcoding (O'Malley et al., 2005). There can also be disagreement in EHR systems if values exist multiple times.

Goldstein et al. (2017) pointed out several advantages and challenges arising from the use of EHR data for predictive modelling. EHR systems represent patient histories over a longitudinal timespan and provide big sample sizes and predictors from different sources for modelling. Hospital networks from a single region can include the majority of the population, and therefore EHR data can be more reflective of the real world than data from prospective cohort studies, which often rely on voluntary participation.

When it comes to deployment, models trained on EHR data can be implemented more easily in HIS and do not have to be translated first.

However, EHR data may suffer from insufficient data quality, represent sicker people on average and underlie informative presence, i.e. the potential information in the presence or absence of patient data (Goldstein et al., 2017). It is still not clear how these factors influence the development and deployment of prediction models, and which biases may occur. Poor EHR data quality cannot only be present in predictors of a model, model accuracy can also be reduced if the predicted outcome is misclassified (Wang et al., 2016). If an EHR system does not fully capture the predicted clinical event, some patients are incorrectly classified as controls during training. Thus, false negative cases will be increased during modelling and this might lead to biased results.

2.2 FROM DEVELOPMENT TO DEPLOYMENT

A recent study by Lee et al. (2020) reviewed predictive models integrated into EHR systems and deployed in clinical practice. Most models were based on linear or logistic regression methods, and the authors concluded that there is still a gap in the evaluation of machine learning models concerning clinical implementation and outcomes.

This section discusses barriers to the implementation of machine learning models, key aspects for a successful adoption by healthcare professionals and ways to enable an interpretation of machine learning models.

2.2.1 *Deployment Barriers of Machine Learning Models*

The integration of machine learning models into clinical routine requires overcoming several obstacles. These obstacles include general barriers to the deployment of decision support systems, such as the technical integration into EHR systems and financial costs (Watson et al., 2020). Various other barriers have been identified which limit the implementation of machine learning models in particular.

First, clear regulations and standards are still missing on how to assess safety and efficacy of machine learning models (Jiang et al., 2017). Without definitions for evaluation, possible stakeholders like hospitals have difficulty in distinguishing adequate models from non-adequate models. At the same time, regulatory requirements can also present implementation barriers if regulations for medical devices apply (Vollmer et al., 2020). Until today, only slightly more than 60 machine learning and AI-based algorithms have been approved as medical devices by the US Food and Drug Ad-

ministration (FDA), most of them for analysing CT, X-ray or MRI images (Benjamens, Dhunnoo, & Meskó, 2020).

Second, challenges for deployment refer to data quality and data transfer. Information systems often lack interoperability, which limits the exchange of data between them. For many developments, after a first successful implementation, continuous supply of data from the HIS is not further provided for the developers. This limits the improvement and further development of already implemented models (Jiang et al., 2017).

Retrospective data used for modelling differ from prospective data in various ways, e.g. in the availability of data at certain time points. Thus, when training prediction models on retrospective clinical data, the model performance can be worse for prospective data (Amarasingham et al., 2014).

Third, several ethical challenges have to be addressed (Char, Shah, & Magnus, 2018). Algorithms based on machine learning are known to be vulnerable to mirror biases of human decision making such as racial discrimination. Biases can also occur if algorithms are trained on different populations than finally implemented or used for. Predictors based on ethnicity, for instance, should be added only with care and for certain reasons (Gijsberts et al., 2015).

In addition, the intent behind a prediction model has to be taken into account. An algorithm that is mathematically optimal is not necessarily ethically optimal (Magrabi et al., 2019). Algorithms can lead to clinical actions that improve quality metrics or reduce healthcare costs, but might not lead to better care or to any other benefit for patients.

Finally, each step of the development and implementation of machine learning models needs to comply with privacy protection and might require patient consent, for example when the need for an intervention is predicted (Amarasingham et al., 2014).

2.2.2 *Adoption of Machine Learning Models*

As few machine learning models have been integrated into clinical routine, there is little research regarding the adoption of machine learning models by healthcare professionals (Amarasingham et al., 2014; Islam et al., 2018).

For a successful adoption in healthcare, machine learning models need to prove their benefit in clinical routine. Prediction models should support or optimise clinical decision making processes in a meaningful way (Vollmer et al., 2020). Predictions of the models should be of high clinical utility and clinicians should be able to take actions and modify a predicted risk (Watson et al., 2020).

Besides providing a benefit, prediction models need to be trustworthy. Spiegelhalter (2020) distinguishes between two types of trustworthiness: (a) trustworthiness of claims made about the system, and (b) trustworthiness of claims made by the system.

Trustworthiness about the system is highly influenced by the utility of an algorithm and its benefit in actual use. Trustworthiness is established when providing information about what an algorithm is able to predict and how it has been evaluated. Communication should include information regarding responsibilities of the development, accuracy of the algorithm, auditability by third parties and fairness to demographic characteristics.

Clinicians are used to interpreting metrics like sensitivity or specificity of diagnostic tests, but are often not comfortable with the interpretation of ROC (receiver operating characteristic) curves or confusion matrices (Watson et al., 2020). This needs to be addressed when communicating the performance of prediction models to healthcare professionals.

Magrabi et al. (2019) also highlight the importance of the correct communication of evaluation results. Results should be understandable to clinicians and managers, especially when it comes to transparency considering limitations and ethical aspects of machine learning models. Healthcare professionals should understand how a system was created and how it impacts their work in order to use it responsibly.

Although trustworthiness demands transparency, the transparency of complex systems does not necessarily provide explainability (Spiegelhalter, 2020).

The trustworthiness of claims made by the system is highly related to the explainability of its claims. Users will trust a prediction if they understand the behaviour of the model and if explanations of predictions are faithful and intelligible (Ribeiro, Singh, & Guestrin, 2016). However, such explanations present big challenges for machine learning models that use many different predictors and that model complex relationships.

2.2.3 *Explainability of Machine Learning-Based Risk Predictions*

Depending on the statistical method used for modelling, some models are more easily interpretable than others. When thinking of model interpretability as a dimension, fully human-guided methods are located on one end of the scale, as they are highly interpretable, whereas fully machine-guided methods on the other end, as they are perceived as black boxes (Beam & Kohane, 2018). While single decision trees are located at the highest level of interpretability, linear combinations of trees such as random forest models are located more in the middle of the scale (Hastie, Tibshirani,

& Friedman, 2009). Deep learning methods are located at the end of the dimension with lowest level of interpretability. They require enormous amounts of data but less human guidance to be trained.

Regardless of the method used for modelling, input variables have different influence on the outcome (Hastie, Tibshirani, & Friedman, 2009). One way to improve the interpretability of models is to provide users with lists of variables ranked by their relative weights of contribution or importance to the prediction (Ribeiro, Singh, & Guestrin, 2016). The most common measure for variable importance in random forest was proposed by Breiman (2001a). The goal of the method is to measure the prediction strength of each predictor, which can then be ranked according to its importance for the model as a whole.

Due to Spiegelhalter (2020), at least the developers of algorithms should be able to give explanations on risk predictions if the algorithm itself is difficult to interpret. However, it remains a duty for developers to partially or fully open the black box of machine learning-based prediction models, so that clinicians can better judge the outcome. There are big difficulties drawing causal inference from predictors of machine learning (Shiffrin, 2016), and this lack of causality presents a concern regarding explainability (Bhatt et al., 2020); variable importance methods are an important step towards the opening of black box models.

Besides improving the trustworthiness of a system, an increase in explainability can reduce the automation bias. Humans assisted by decision support systems tend to over-rely on them and stop questioning the output (Magrabi et al., 2019). Improving the interpretability of a machine learning model and explaining the underlying predictors can lower this bias.

2.3 EVALUATION OF MACHINE LEARNING MODELS IN HEALTHCARE

Although many machine learning models achieve good performance in test and validation data, there are hardly any follow-up studies on the performance of such models in dynamic decision-making situations (Islam et al., 2018). Few machine learning-based risk prediction models can be found in late-stage clinical development, and there has been little prospective validation determining their benefits and usefulness (Topol, 2019). As a result, little is known about how healthcare professionals interact with machine learning models and how they perceive them.

2.3.1 *Considerations for Evaluation*

Magrabi et al. (2019) presented key considerations and practical aspects for the evaluation of machine learning-based clinical decision support systems. Although many of their suggestions refer to the evaluation of models before an implementation, they also highlight several points of interest after it.

During the development of a model, appropriate performance measures need to be chosen depending on the specific use case. While an algorithm for triage tries to achieve highest discrimination, an algorithm predicting risk of complication or mortality needs to be accurate for all patients. At this point, possible differences between training cohorts and patient populations for prospective predictions should be evaluated.

Prediction models should further be evaluated considering their risk of data quality issues and their generalisability to new situations. It has to be assessed whether the data needed by the model are available in the setting selected for implementation. This aspect is related to the data quality constraints discussed in Section 2.1.3.

After a stable algorithm is obtained, evaluation should compare the decision making of healthcare professionals with and without the decision support system. The idea of comparing the performance between humans and machine leads back to an experiment in the 50s, the Turing test (Turing, 1950). This comparison is further related to the validation of utility; although a model predicts a risk accurately, it might not be meaningful in clinical practice, not fit into clinical workflows and thus not give any support to healthcare professionals or benefit for patients.

Magrabi et al. (2019) further illustrate the need of an ongoing surveillance in order to monitor changes after the deployment of machine learning algorithms. Different patient populations, treatment possibilities and organisational and social impacts can change the data used for prediction. Thus, the need for recalibration of the models should be assessed regularly after the deployment.

2.3.2 *Evaluation of Machine Learning Acceptance*

User satisfaction and the acceptance of a system play a key role when evaluating any decision support system. Wyatt & Spiegelhalter (1990) stated already in the early nineties that "[...] medical expert systems will not succeed unless they are wanted, are usable in the clinical environment and draw conclusions that seem reasonable to the user".

Two recent studies reported on the acceptance of machine learning-based risk prediction by clinicians. The first study by Brennan et al. (2019) evaluated the application MySurgeryRisk in a prospective pilot study. MySurgeryRisk is an automated analytics framework, which provides predictions for several postoperative complications based on clinical data from the EHR system using generalised additive models and random forest models (Bihorac et al., 2019).

Brennan et al. (2019) assessed the usability and the performance of the framework before its actual implementation. 150 patient cases were simulated in order to study the interaction between physicians and the algorithm. 20 surgical intensivists provided risk ratings on a scale from 0% to 100% for all postoperative complications before and after seeing the scores of the algorithm. For around 75% of the cases, physicians changed their risk predictions after interacting with the algorithm. However, only five out of ten physicians who participated in the usability survey reported that the algorithm helped them in their decision-making process. The same amount reported that they would use the application for counselling patients preoperatively, and eight of ten found it easy to use.

The second acceptance study was conducted by Ginestra et al. (2019). The authors assessed the clinical perception of the Early Warning System (EWS) 2.0 predicting sepsis in non-intensive care unit (ICU) inpatients (Giannini et al., 2019). EWS 2.0 uses a random forest classifier to predict severe sepsis or septic shock, and recalculates the risk hourly with the latest EHR data.

During a pilot study of 14 months, predictive alerts were sent to the care team (Ginestra et al., 2019). Users were questioned twice for a total of 362 triggered alerts over six weeks. 287 nurses and physicians completed a first survey six hours after an alert, resulting in a response rate between 30-50%. 47 participants completed the second survey 48 hours after the alert (response rate 24-41%). The majority of the users reported no changes in their perception of a patient's sepsis risk, and the minority identified new clinical findings or reported changes in risk management. Nurses found the alert more helpful than physicians and reported more often that the alerts improved care. Physicians mainly criticised that the system fired too late, and that it triggered mostly for already known abnormalities. They further requested an improvement of transparency of predictors that led to an alert trigger.

2.4 THE USE CASE DELIRIUM

This section illustrates the need and usefulness of accurate risk prediction of delirium, and the benefit of supporting its prediction using machine learning methods.

2.4.1 *Clinical Perspectives of Delirium*

Delirium is a syndrome of acute confusional state and is characterised by an acute decline in cognition and attention (Inouye, 2006). The onset of delirium is multifactorial, depending on predisposing factors and precipitating factors. Especially elderly patients are often affected by delirium and suffer from its consequences.

Delirium patients have longer hospital stays (McCusker et al., 2003) and their health is often affected by massive complications. They show reduced cognitive rehabilitation (Girard et al., 2010) and have an increased need for long-term care (Bickel et al., 2008). Mortality rates are increased for delirium patients not only in the short term (Inouye, Westendorp, & Saczynski, 2014; Lin et al., 2004) but also in the long term, together with an increased risk of institutionalisation (Eeles et al., 2010; Witlox et al., 2010).

Different subtypes of delirium have been recognized, depending on the psychomotor behaviour of a patient: While *hypoactive* delirium is characterized by unresponsiveness or slowing movement, symptoms of *hyperactive* delirium range from simple restlessness to agitation. Mixed forms of delirium include symptoms of both types (Yang et al., 2009; de Rooij et al., 2005). Depending on these subtypes of delirium, different patient management and treatment strategies are necessary. Hyperactive patients attract more attention and are thus easier to detect than hypoactive patients (de Rooij et al., 2005), but their treatment presents a higher burden for nurses (Schmitt et al., 2019).

Occurrence rates of delirium range widely between studies, depending highly on the population studied. In a systematic review from Inouye, Westendorp, & Saczynski (2014), prevalence rates ranged from 18 to 35% and incidences from 11 to 14% in general medical departments. A review from Siddiqi, House, & Holmes (2006) found prevalence rates ranging from 10 to 31% and incidence rates from 3 to 29% for medical inpatients. Even higher incidence rates of delirium can be found in ICU and in postoperative settings (Inouye, Westendorp, & Saczynski, 2014).

However, in many cases delirium is overlooked and remains undiagnosed (Lange et al., 2019). The underdetection of delirium and a delayed treatment of the syndrome can lead to complications and can further delay the treatment of the disease a patient has been primarily admitted for. Due to increased monitoring and the commitment of more staff members, higher nursing time is necessary for delirium patients (Weinrebe et al., 2016). Hence, delirium results not only in a burden for patients and their families (Schmitt et al., 2019), but also in substantially higher healthcare costs (Leslie, 2008). As a conclusion, patients, families, healthcare professionals and hospital networks benefit from a successful prevention of delirium.

2.4.2 *Risk Factors and Prevention of Delirium*

The leading predisposing risk factors identified for delirium are dementia, cognitive or visual impairment, history of alcohol misuse and age older than 70 (Inouye, Westendorp, & Saczynski, 2014). Precipitating risk factors that can trigger the onset of a delirium include sedative-hypnotic drugs, use of urinary catheter, infection or surgery. In addition, several environmental risk factors have been identified that influence symptoms of inpatients with delirium (McCusker et al., 2001). A higher number of room changes, long-term care, absence of a clock in the room or absence of reading glasses were associated with worse outcomes; frequent presence of family members was related to better outcomes.

There is high evidence that the occurrence of delirium can be prevented in many cases (Hshieh et al., 2015; Inouye et al., 1999). In accordance with the multifactorial aetiology of delirium, multicomponent approaches for prevention and treatment were found to be most effective (Inouye, Westendorp, & Saczynski, 2014). One example is the use of clinical protocols to detect risk factors (Young et al., 2008). Such protocols commonly include checks for urinary tract infection, dehydration, oral hygiene, disorientation, vision and hearing or sleep quality. Detection and further modification of these risk factors can reduce delirium by up to one third.

Concordantly, a Cochrane review from Siddiqi et al. (2016) found strong evidence for effective prevention of delirium in hospitalised patients using non-pharmacological multicomponent interventions. Although guidelines recommend the use of antipsychotic medication to lower severity of delirium, there is no convincing evidence for a pharmacological prevention or treatment (Inouye, Westendorp, & Saczynski, 2014; Siddiqi et al., 2016; Young et al., 2008).

2.4.3 *Screening and Diagnosis - State of the Art Methods*

It seems inevitable to detect patients with highest risk as early as possible in order to successfully prevent the onset or mitigate signs and symptoms of delirium. Even though several screening methods have been established, they have their limitations when it comes to delirium management in hospitals.

One of the most common tools to detect delirium is the Confusion Assessment Method (CAM). It was developed by Inouye et al. (1990) based on the diagnostic criteria from the Diagnostic and Statistical Manual of Mental Disorders, Version 3, (DSM-3) of the American Psychiatric Association (1980). In accordance with the DSM-3, CAM assesses (1) an acute change in mental status with a fluctuating course, (2)

inattention, (3) disorganized thinking, and (4) altered level of consciousness. CAM is considered to be positive (and indicates delirium) with the presence of (1) and (2), and either (3) or (4).

Another common screening tool is the Delirium Observation Screening (DOS) scale developed by Schuurmans, Shortridge-Baggett, & Duursma (2003). The aim of the tool is to facilitate early recognition of delirium. Just like CAM, it is based on the DSM criteria of delirium. However, DOS consists of 13 items which are more specific than the items of CAM, e.g. including the question whether a patient "knows which part of the day it is". It is recommended to apply the DOS scale three times a day during regular care. Even though some of the items are fast to answer ("pulls IV tubes, feeding tubes, catheters etc."), other items require more detailed observations ("reacts slowly to instructions").

In hospital routine, screening scales like CAM or DOS are often time-consuming for healthcare professionals, especially when considering frequent changes in rosters and high numbers of patients. As both scales were developed mainly for screening and detection, their items are based on diagnostic criteria and assess signs and indicators for delirium that are already present. When using exclusively such items for delirium risk prediction, patients with several other predisposing and precipitating risk factors may go undetected.

In the International Classification of Diseases tenth revision (ICD-10), delirium has the codes in the range of F05.0 to F05.9. While the English version of ICD-10 describes F05 as "delirium due to known physiological conditions", the German modification (ICD-10-GM) describes it as delirium not caused by alcohol or other psychotropic drugs ("Delir, nicht durch Alkohol oder andere psychotrope Substanzen bedingt"). This indicates the fact that delirium is not only common in elderly people with predisposing and precipitating risk factors, but can also occur in younger patients along with alcohol withdrawal.

2.4.4 *Alcohol Withdrawal with Delirium*

Prevalence of alcohol dependence is around 9% in Europe, and out of all Europeans with alcohol dependence up to 22% receive medical treatment (Rehm et al., 2015). When reducing or discontinuing alcohol consumption, about 50% of patients with alcohol-use disorders show signs of withdrawal (Schuckit, 2014).

The ICD-10 code F10.4 describes a "withdrawal state with delirium", colloquially called delirium tremens. The incidence of delirium in patients with alcohol withdrawal varies from 3 to 20% (Lee et al., 2005; Mainerova et al., 2015). Delirium with alcohol

withdrawal begins around three days after the first withdrawal signs and symptoms, and lasts about 2 or 3 days. Up to 4% of the patients with withdrawal delirium die because of seizures, hyperthermia or cardiac arrhythmias, but death can be prevented with early diagnosis and the right treatment (Schuckit, 2014).

Although this kind of delirium is excluded in most prediction models for delirium, it is equally important to identify patients at high risk. Possible risk factors for alcohol withdrawal delirium include previous episodes of delirium or seizures during withdrawal, infectious diseases or hypertension (Mainerova et al., 2015). In addition, several laboratory values and in particular elevated liver enzymes indicate a higher risk, including alanine aminotransferase, aspartate aminotransferase and gamma-glutamyltransferase, or mean corpuscular volume and carbohydrate-deficient transferrin.

2.5 CLINICAL PREDICTION MODELS FOR DELIRIUM

2.5.1 *Published Reviews on Delirium Prediction Models*

Several reviews on prediction models for delirium have been published in recent years. van Meenen et al. (2014) reviewed models for postoperative delirium and conducted a meta-analysis. Although some prediction models had been internally and externally validated, the authors did not recommend any of them for clinical practice as they considered the evidence for good performance too weak. Further limitations were a lack of generalisability to other populations and the frequent use of age and the Mini-Mental-State-Examination score as predictors. While the authors criticized the predictor age to be not differentiating well in patients above 70 years, the Mini-Mental-State-Examination requires training and is time consuming.

Newman, O'Dwyer, & Rosenthal (2015) conducted a review of delirium prediction models for hospitalised internal medicine patients excluding ICU and post-surgical patients. None of the models included in their review had been adopted to clinical practice, and the authors mentioned several challenges for implementation: Some models included variables which were difficult to measure on admission, and all of them required at least some values to be completed by hospital staff and were thus found to be time consuming.

Another review with importance to this thesis was conducted by Lindroth et al. (2018). It included prediction models developed on inpatients older than 60 years, and excluded alcohol-related delirium. The authors pointed out the need for more dynamic models developed with other statistical methods including machine learning.

Furthermore, they criticized the clinical utility if only patients at highest risk are detected, while prediction for patients of moderate or ambiguous risk might achieve a greater clinical value.

2.5.2 *Prediction Models with Importance to this Thesis*

For this thesis, a systematic review on the discriminative performance of prediction models for delirium using EHR data was conducted. Details of the search strategy and a PRISMA flow diagram (Liberati et al., 2009) of the reviewed research articles are included in the Appendix in Table B.1 and Fig. A.1, respectively. Eight publications were identified by the systematic review; three of them were excluded as they had been prepared by our research group and describe the prediction models evaluated in this thesis (Kramer et al., 2017; Veeranki et al., 2019; Veeranki et al., 2018). In addition, the study from Kim et al. (2016), which was included in the review by Lindroth et al. (2018), was added to the literature overview of this section.

Table 2.1 presents the results of the six reviewed studies. Besides the modelling method and discriminative performance, numbers of patients included in the training and the final numbers of model predictors are illustrated.

Table 2.1: Reviewed EHR-based models predicting delirium in inpatients.

Publication	Method	N[a]	Pred.[b]	Results			
				AUROC [95%-CI]		Sens.	Spec.
Rudolph et al. (2016)	G	-	6	0.69	[0.61–0.77]	64.0%	-
de Wit et al. (2016)	LR	1,291	12	0.78	[0.74–0.81]	78.2%	63.5%
Kim et al. (2016)	LR	561	9	0.94	[0.91–0.97]	80.8%	92.5%
Halladay et al. (2018)	RF	27,625	16	0.91	[0.90–0.92]	-	-
Corradi et al. (2018)	RF	51,240	128	0.86	[0.84–0.88]	-	-
Wong et al. (2018)	GBM	14,227	796	0.86	-	59.7%	90.0%

Note: [a]Number of patients included for training; [b]Number of model predictors; G: Guideline based; LR: Logistic regression; RF: Random forest; GBM: Gradient boosting machine

Rudolph et al. (2016) weighted delirium risk factors identified by the National Institute for Health and Clinical Excellence (NICE) (2010) in the UK and included them in the e-NICE algorithm. Six risk factors for delirium were included in the algorithm based on the meta-analysis: impaired cognition, impaired vision, severity of illness, fracture, infection, and age. The e-NICE algorithm achieved an area under the receiver operating characteristic curve (AUROC, described in Section 3.1.1.2) of 0.69 in

a prospective validation cohort including 246 patients. The performance of the model increased to 0.74 when applying additional cognitive screening.

The prediction model developed by de Wit et al. (2016) used 12 predictors derived from univariable regression analyses including age, prescribed medication and laboratory values. The final logistic regression model achieved an AUROC of 0.78 in a cohort of patients older than 60.

Kim et al. (2016) developed the Delphi score, which was also the best discriminating model in the review by Lindroth et al. (2018). The Delphi score achieved an AUROC of 0.94 using nine variables for prediction. However, some of the variables need to be assessed during or after surgery (e.g. ICU admission) which makes the model unsuitable for early risk prediction. Lindroth et al. (2018) also criticized the results of the Delphi model because some variables for prediction were collected after the possible onset of delirium. The data collection for the model occurred within the first 24 hours after surgery, while delirium assessment began immediately after surgery. Due to this data overlap, the performance of the score might be exaggerated. Furthermore, the score is only applicable for patients older than 60 who underwent major general surgery (gastrointestinal, hepatobiliary-pancreatic, colorectal, vascular, or trauma surgery).

The model proposed by Halladay, Sillner, & Rudolph (2018) achieved an AUROC of 0.91 when predicting delirium prevalent at admission. Similar to Rudolph et al. (2016), the model was developed based on the NICE guideline but in combination with random forest. Their definition of delirium was very broad and the study population included US veterans only. The use of this model can rather help not to overlook a present delirium than prevent the occurrence of delirium.

Two of the identified publications are highly relevant for this thesis, the study by Corradi et al. (2018) and the study by Wong et al. (2018).

Corradi et al. (2018) developed a random forest model predicting delirium in hospitalised patients. The predicted outcome was a positive CAM assessment at least 48 hours after admission. EHR data used for prediction included demographic data, comorbidities using ICD codes, vital signs, medication orders, procedures and predictors from the Richmond Agitation-Sedation Scale. Only data available during the first hours of hospital stay were included for modelling. Using 128 predictors, the model achieved an AUROC of 0.861 on validation data of 10,000 non-ICU patients.

Wong et al. (2018) compared five different machine learning methods for predicting delirium at hospital admission: penalized logistic regression, GBM, artificial neural network with a single hidden layer, support vector machine (SVM), and random forest. The models were trained on non-ICU patients without delirium present on admission

and included 796 predictors from EHR data such as admission diagnoses, medications, laboratory values, vital signs, demographic data and nursing data. The best performing model, a GBM model, achieved an AUROC of 0.855 on validation data of 3,996 patients. The sensitivity of the model was 59.7% at a specificity set at 90%.

2.5.3 *Summary*

Several prediction models have been published using EHR data for delirium prediction, but only few of them used machine learning methods for modelling. Several aspects limit the deployment of the prediction models in clinical practice. Models were developed on narrowly defined cohorts and are thus not generalisable to other populations, data need to be completed by hospital staff or values are not available at point of admission. Thus, most models will not be able to support an early risk prediction of delirium with great clinical value.

However, the biggest limitation found is missing evidence from the performance of the models in clinical practice (Newman, O'Dwyer, & Rosenthal, 2015). Although some models performed very well in retrospective data sets, none of the reviewed machine learning models have been implemented and prospectively evaluated in a routine clinical setting.

3

METHOD

This chapter describes the methods used for the evaluation of the delirium risk stratification tool. First, the most important performance measures for prediction models are introduced. Second, the expert group is described that was involved in the pilot study. Third, the methods used to identify patients with delirium are presented. Finally, the study design is presented answering the three research questions raised in Chapter 1.

Specific methods and details of the study design for each research question are described in the corresponding chapters (Chapter 5, Chapter 6 and Chapter 7). All analyses for this thesis were performed using the R software in Version 3.4 and 3.6.

3.1 MEASURES OF PERFORMANCE

The assessment of the performance of a classification model is important during its development (e.g. when comparing different models) and during its evaluation (e.g. to determine whether it performs sufficiently for the defined use) (Hand, 2012).

A major contribution to the assessment of the performance of prediction models was made by Steyerberg et al. (2010). According to the authors, two main characteristics need to be addressed when evaluating prediction models with binary outcomes: discrimination and calibration. This section presents common measures for both characteristics.

3.1.1 *Measures of Discrimination*

The main goal of a classification model is to select the best available threshold to discriminate between two or more classes (Hand, 2012). The threshold depends on the circumstances of deployment, and it is often let unspecified during model development. However, as soon as a prediction model is applied, the threshold has to be set. Therefore, different measures of performance are needed when evaluating prediction models with or without given thresholds.

3.1.1.1 *Measures with Specified Threshold*

Most classification problems in the medical domain are evaluated using thresholds. Based on the thresholds, a confusion matrix can be calculated comparing the actual outcome to the predicted outcome. For binary classifiers, this results in a 2x2 matrix, from which several measures of discrimination can be reported for prediction models.

- *Sensitivity*, also called *true positive rate* or *recall*, is the number of correctly predicted positive cases among the total number of positive cases.
- *Specificity*, or *true negative rate*, is the number of correctly predicted negative cases among the total number of negative cases.
- *False positive rate* is the number of false positive cases among all negative cases, or 1 - Specificity.
- *False negative rate* is the number of false negative cases among all positive cases, or 1 - Sensitivity.
- *Positive predictive value*, also called *precision*, is the number of the correctly predicted positive cases among all positive predicted cases.
- *Negative predictive value* is the number of correctly predicted negative cases among all negative predicted cases.

3.1.1.2 *Measures Without Specified Threshold*

Without a specified threshold, measures derived from a confusion matrix cannot be used. One way to overcome the constrain of an unspecified threshold is to use multiple possible threshold values to describe the model performance (Hand, 2012).

A common measure of discrimination with unspecified threshold is the *area under the receiver operating characteristic curve* (AUROC), which is equal to the *concordance statistic* (*c*) for binary outcomes (Steyerberg et al., 2010). Hosmer, Lemeshow, & Sturdivant (2013) provided a suggestion for the interpretation of AUROC values describing discriminative performance (illustrated in Table 3.1).

In this thesis, AUROC represents the probability that a randomly chosen patient of the group of patients without delirium has a lower predicted probability than a randomly chosen patient of the group of patients with delirium.

When reporting AUROC as a measure of performance, the metric is usually visualized plotting the curve of the *receiver operating characteristic* (ROC). In ROC plots, the x-axis shows the false positive rate or 1 - specificity, and the y-axis the corresponding true positive rate or sensitivity.

Various methods have been proposed for confidence intervals (CI) of ROC curves. For this thesis, the 95%-CIs of ROC curves were computed based on the DeLong

method (DeLong, DeLong, & Clarke-Pearson, 1988) using the *pROC* package (Robin et al., 2011) in R.

Table 3.1: Interpretation of AUROC values according to Hosmer, Lemeshow, & Sturdivant (2013)

Value	**Interpretation**
AUROC = 0.5	No discrimination (equal to flipping a coin)
$0.5 < \text{AUROC} < 0.7$	Poor discrimination (not much better than coin toss)
$0.7 \leq \text{AUROC} < 0.8$	Acceptable discrimination
$0.8 \leq \text{AUROC} < 0.9$	Excellent discrimination
$\text{AUROC} \geq 0.9$	Outstanding discrimination

3.1.1.3 *Choosing the Right Evaluation Measures*

As outlined by Magrabi et al. (2019), different performance measures are necessary depending on the use case of a prediction model. While some prediction models aim for a high specificity, in other scenarios high sensitivity is preferred.

The selection of measures depends not only on the use case, but also on the underlying data. AUROC values with 95%-CI and corresponding ROC plots serve as an overall measure of discriminative performance.

There is evidence that for imbalanced data, precision-recall plots can be more informative than ROC plots (Saito & Rehmsmeier, 2015). In the original EHR data used for the development of the prediction models evaluated in this thesis, the number of delirium patients and non-delirium patients was highly imbalanced. However, resampling methods were used during the development (see Section 4.1.1), and, hence, ROC plots were considered to be of sufficient information.

For the deployed delirium risk stratification tool, two thresholds were set in order to stratify patients into three risk groups. As illustrated in Section 4.2.1, thresholds varied between different clinical department. In order to calculate performance measures with a specified threshold, the lower threshold was used to separate the predicted risk groups into a *low risk* group and a *high* or *very high* risk group.

3.1.2 *Measures of Calibration*

Although measures of calibration are recommended by the TRIPOD guideline for transparent reporting of multivariable prediction models (Collins et al., 2015; Moons

et al., 2015), they are less frequently assessed than measures of discrimination (Van Calster et al., 2016). Measures of calibration demonstrate the agreement between the predictions of a model and observed outcomes. In other words, calibration refers to the reliability of probabilistic claims.

One method to examine the calibration of a model are calibration plots (Steyerberg et al., 2010). On the x-axis, predictions of an algorithm are presented, while the y-axis shows the observed outcome. Perfect predictions should be located at the 45 degree line. For binary outcomes, the x-values often present percentiles of the predicted risk probabilities, and the y-values present relative frequencies of the outcome.

Two main aspects show the importance of measuring calibration in this thesis. First, there is evidence that an algorithm developed on data with high incidence of the predicted outcome might systematically overestimate the risk when used in a setting with low incidence (Van Calster et al., 2019). As resampling methods led to a higher incidence of delirium in the training data (see Section 4.1.1), measures of calibration will provide insights on a possible overestimation.

Second, an underestimation of the predicted risk can increase the number of false negative cases because patients with delirium would then be stratified to a low risk group instead of to a high risk group. This might further result in a low discriminative performance, a poor acceptance by healthcare professionals, and an undertreatment of patients (Van Calster et al., 2019).

3.2 EXPERT GROUP

Clinical experts were constantly involved in the evaluation of the delirium risk stratification tool. An expert group was set up in order to enhance participation of healthcare professionals, define user requirements and consider clinical experience during the whole implementation and evaluation period.

The evaluation of the tool was carried out in cooperation with KAGes. KAGes is the public health care provider in Styria, a federal state of Austria, and runs several subsidiary hospitals. One of them is the federal state hospital Landeskrankenhaus (LKH) Graz II, which served as a pilot site.

Staff members from the pilot hospital were nominated for participation in the expert group by the heads of the surgical and internal medicine departments. Depending on the clinical roster, up to five senior physicians and five ward nurses attended the meetings. Physicians and nurses also contributed to the evaluation of the acceptance of the delirium risk stratification tool.

The expert group was involved in decisions regarding the visualization in the HIS and determining the number of high-risk-patients to be visualized. Available resources for prevention were analysed in the departments, and thresholds for the three risk groups were adapted in agreement with the clinical experts (see Section 4.2.1).

During the pilot phase, the expert group facilitated usability engineering and improvement of the delirium risk stratification algorithm. In addition to clinicians, five machine learning engineers and IT professionals in charge of the maintenance of the HIS were included as experts. Of the 15 expert group members eight were male (53%) and seven female (47%).

3.3 IDENTIFICATION OF PATIENTS WITH DELIRIUM

A task with high importance for the evaluation was the identification of patients with delirium in the EHR data. Two methods were used for identification during the evaluation, which were identical for model development. First, EHR data of patients were searched for delirium-related ICD-10 codes, and second, discharge summaries were screened using text mining methods.

EHR data of a patient's hospital stay were searched for the ICD-10 codes F05 (including all subcategories) and F10.4 if alcohol withdrawal was present. Whenever patient transfers occur within the KAGes network, data are documented using different identification numbers for the stay in each hospital. Thus, for each patient, EHR data from the KAGes network were screened up to 14 days after discharge of the primary hospitalisation under consideration.

As physicians are not always familiar with ICD-10 codes, delirium is not always correctly coded, e.g. with the too unspecific symptom code R41.0 (Disorientation, unspecified). ICD-10 coding is mostly required for hospital administration, and mentions of diseases and clinical outcomes are generally more accurate in clinical texts of EHR systems.

Therefore, as a second method of identification, discharge summaries were screened for delirium. Discharge summaries from the hospitalisations and up to 14 days after discharge were exported. For text mining in R, approximate string matching with Levenstein distance measures depending on the length of the searched strings was used. Search words included the following delirium related words in German: *Delir* (delirium), *akute Verwirrtheit* (sudden confusion), *passagerer Verwirrtheitszustand* (transitory confusional state), *Durchgangssyndrom* (transitory syndrome). The corresponding paragraph of the discharge summary was extracted for all matching strings and was

checked manually. Patients with a delirium verified were included in the group of delirium patients.

3.4 STUDY DESIGN

Fig. 3.1 illustrates the timeline for the study conducted in this thesis.

The implementation process started in November 2017, when a first expert group meeting took place. The goal of the meeting was to communicate the aims of the expert group meetings, to inform clinical experts about strengths and weaknesses of the tool, and to receive a first feedback concerning the acceptance.

In February and March 2018, training sessions were offered to healthcare professionals from the participating wards. The main objective was to stimulate the understanding and the adoption of the delirium risk stratification tool. The tool was implemented in April 2018 in the pilot site LKH Graz II (Hospital I). The use of it was voluntary for all healthcare professionals.

The following sections provide an overview of the methods used to fulfil the aims of the three parts of this thesis:

1. the evaluation of the performance of the delirium risk stratification algorithm in a prospective setting
2. the evaluation of the technology acceptance of the tool
3. the evaluation of the long-term performance in a multicentre setting

3.4.1 *First Part: Evaluation of the Performance in a Prospective Setting*

In the first part of the thesis, illustrated in orange in Fig. 3.1, the performance of the delirium risk stratification algorithm was evaluated in a clinical setting in Hospital 1. The prospective evaluation of the performance was split into two parts.

First, prospective predictions of the algorithm were analysed to determine the performance in the clinical setting. The evaluation of prospective predictions was carried out from 1^{st} of June until 31^{st} of December 2018.

Second, comparisons were conducted between the algorithm and clinical experts for a sample of general internal medicine and gastroenterology patients. A protocol was developed for the clinical assessment of delirium risk. It was completed by experienced ward nurses within the first 24 hours of a patient's hospital stay. The risk rating by the nurses was then compared to the risk prediction of the algorithm.

The first comparison (Comparison 1) was conducted in February 2018 before the prediction was visible in the user interface of the HIS, and was thus blinded. This

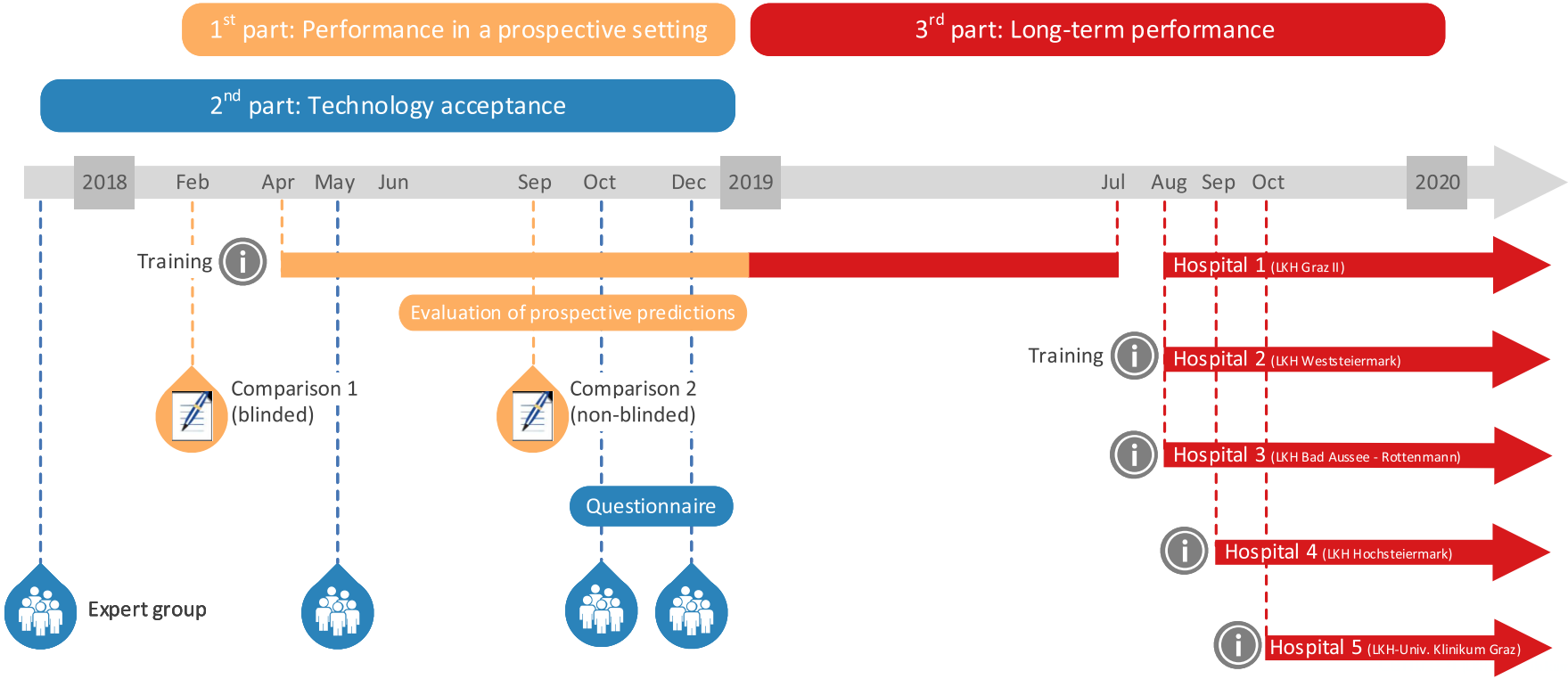


Figure 3.1: Timeline for the three parts of this thesis.

comparison provided a first quality assessment of the algorithm's accuracy in clinical routine. The second comparison (Comparison 2), was carried out in September 2018 and was non-blinded.

3.4.2 *Second Part: Evaluation of the Technology Acceptance*

The second part of the thesis evaluated the technology acceptance of the delirium risk stratification tool. The goal was to obtain a detailed picture of the deployment of the tool in a clinical setting.

A convergent parallel design for a mixed methods study combined quantitative and qualitative assessments, illustrated in blue in Fig. 3.1. For the qualitative assessment, opinions and comments from clinical experts were collected during four expert group meetings.

For the quantitative assessment, questionnaires were distributed to healthcare professionals from participating wards at the end of the pilot study. In order to get an impression of different user groups, physicians and nurses of all experience levels were encouraged to voluntarily participate in the quantitative assessment.

3.4.3 *Third Part: Evaluation of the Long-Term Performance*

The third part of the thesis focused on the long-term performance of the delirium risk stratification algorithm in a multicentre setting. Starting from August 2019, the tool was implemented in four other KAGes hospitals in addition to the pilot site LKH Graz II (Hospital 1).

Deployment periods for each hospital are illustrated in red in Fig. 3.1. At the end of the year 2019, data for prospective risk stratification were available from five different hospitals of KAGes over a period of at least three months.

3.4.4 *Ethical Approval*

The study received approval from the Ethics Committee of the Medical University of Graz (30-146 ex 17/18).

4 MATERIALS

In 2018, the delirium risk stratification tool was first implemented in KAGes. The aim of the tool is to support the detection of patients with risk of delirium within the first hours of admission. No recommendation is given by the tool on how to proceed with patients at risk, and the prevention of delirium is thus performed only according to the standards of each ward.

As the tool stratifies hospitalised patients according to their delirium risk into different risk groups, the term *risk stratification* is used in this thesis instead of the more general term *risk prediction*.

This chapter describes the development and functionalities of the delirium risk stratification tool. First, the development of the prediction models is described. Second, the integration of the developed models in the risk stratification algorithm and its use for prospective prediction is elaborated on. Third, the visualization of the predicted risk in the user interface of the HIS is illustrated.

4.1 DEVELOPMENT OF PREDICTION MODELS FOR DELIRIUM

4.1.1 *Modelling Method*

All prediction models described in this thesis were developed according to current recommendations for predictive modelling by Kuhn & Johnson (2013). The *caret* package (Kuhn, 2017) and associated packages were used for the model development in R.

Every data set used for modelling was split into training data (75%) and test data (25%), with an equal distribution of the predicted outcome variable. Models were trained on the training data using a 10-fold cross-validation in order to prevent overfitting. The trained models were tested on the test data, and the best performing model was chosen for further development.

Data sets used for modelling are often imbalanced with a different number of negative controls and positive cases. Because the occurrence of delirium is not always recorded in EHR systems, the number of positive cases of delirium in the data is even

lower than the actual prevalence. Thus, data sets used for modelling are prone to be imbalanced.

One way to overcome this problem is the use of resampling methods such as (a) oversampling or upsampling the smaller class to make it reach the size of the larger class, or (b) downsampling or undersampling to fit the larger class size to the smaller class size (Estabrooks, Jo, & Japkowicz, 2004). In order to select the best performing model, both methods were used with the *caret* package.

4.1.2 *Data Used for Modelling*

For the model development, EHR data of KAGes were used. Due to the coverage of more than 90% of all hospital beds in Styria, KAGes has access to approximately two million longitudinal patient histories.

Routine clinical data of KAGes are stored in openMEDOCS, a HIS based on IS-H/i.s.h.med information systems and implemented on SAP platforms. For every inpatient stay in a KAGes hospital, the following structured information is available:

- demographic data, e.g. age or sex
- transfer data, e.g. ward of admission
- diagnoses, coded with ICD-10
- laboratory data, mapped to international LOINC (Logical observation identifiers names and codes)
- procedures, mapped to Austrian procedure codes
- nursing assessment, e.g. visual impairment and orientation

Besides, further information can be retrieved for each hospitalisation, e.g. information on surgery, ICU documentation, malnutrition screening, etc.

Electronic drug dispensation and drug prescribing has only recently been implemented in KAGes hospitals. Thus, for hospital stays before 2020, information on drugs needs to be retrieved from discharge summaries using text mining.

In addition to the structured information, substantial information of patient histories is included in clinical texts of the EHR system. This mostly unstructured free-texts include discharge summaries, daily nursing notes, or notes from diagnostic procedures, e.g. psychiatric notes and logopaedic notes.

4.1.2.1 *First Model Developments*

In 2016, the first machine learning models predicting delirium were developed (Kramer et al., 2017). The binary outcome was delirium coded during the current hospitalisation (ICD-10 code F05, including all subcategories). In the KAGes network, diagnoses are

often coded in the EHR system close to discharge or up to 14 days after discharge. Therefore, occurrence of delirium was defined as being diagnosed and coded during the hospital stay and up to 14 days after discharge.

EHR data of 8,561 internal medicine patients were extracted for the years 2013 until 2016. The final data set resulted in 858 features including demographic data, previous ICD-10 coded diagnoses, laboratory data, nursing assessment and procedures.

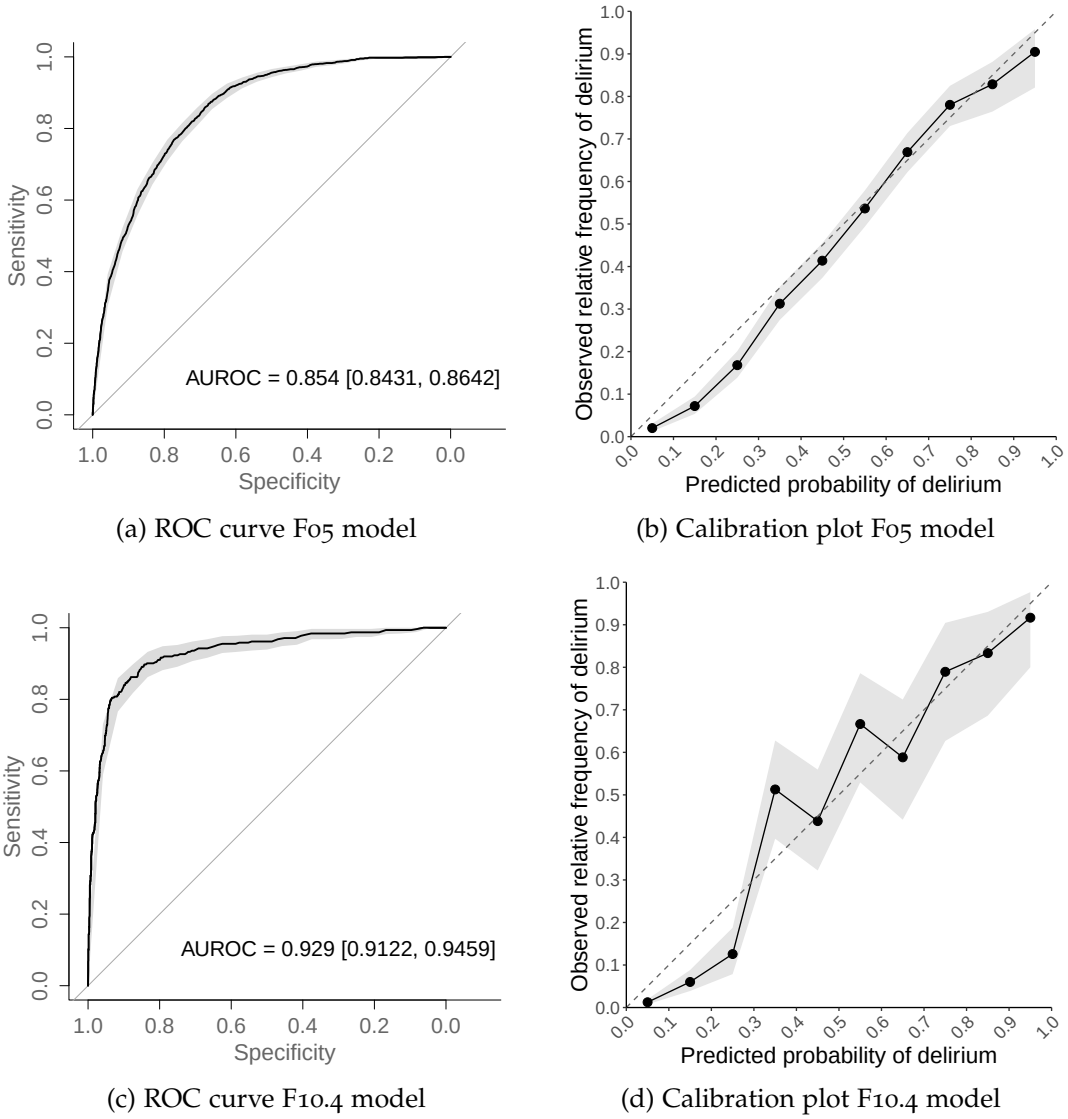
Different machine learning methods were used for modelling including linear discriminant analysis, logistic regression, SVM, k-nearest neighbour, GLM with elastic net regularization (a combination of lasso and ridge regression) and random forest. A comprehensive description of all methods is provided by Hastie, Tibshirani, & Friedman (2009).

The discriminative performance of the trained models in the test data set ranged from acceptable to outstanding. The k-nearest neighbor method achieved the lowest performance with an AUROC of 0.79 [0.774 - 0.815]. The best model was a random forest model, trained with the R package *randomForest* included in the *caret* package. It achieved an outstanding performance with an AUROC of 0.91 [0.897 - 0.922], and slightly outperformed the SVM model with an AUROC of 0.90 [0.891 - 0.917].

4.1.2.2 *Further Development of the Random Forest Model*

During the year 2017, the F05 model developed by Kramer et al. (2017) was adapted as follows:

- In addition to patients from internal medicine departments, patients from surgical departments were added to the cohort.
- Feature selection was performed using a frequency-based approach, resulting in 584 features for modelling. Feature groups and examples are presented in Table B.2 in the Appendix.
- The random forest model was trained and tested on a larger data set with over 19,900 hospitalisations. The data set included 13,445 hospitalisations without delirium (controls) and 6,460 with delirium (cases).
- Discharge summaries of all hospitalisations included in the data set were exported and screened for delirium as described in Section 3.3. Matching strings were checked manually and patients with a delirium verified were included as cases.



(a) ROC curve F05 model

(b) Calibration plot F05 model

(c) ROC curve F10.4 model

(d) Calibration plot F10.4 model

Figure 4.1: Discriminative performance and calibration of the random forest model predicting ICD-10 code F05 trained on 19,905 patients in (a) and (b), and a random forest model predicting ICD-10 code F10.4 trained on 9,872 patients in (c) and (d) (adapted from Jauk et al. (2020))

Although alcohol withdrawal state with delirium (ICD-10 code F10.4) is quite distinct from delirium coded by F05 in terms of aetiology and pathophysiology, clinicians from the expert group (see 3.2) found it crucial to include both predictions because of their similarity in signs, symptoms and consequences. Thus, a model predicting occurrence of delirium coded as F10.4 was trained on a data set of 9,872 hospitalisations and 425 features. Examples of features used for prediction are summarised in Table B.2 in the Appendix.

The discriminative performance and the calibration of the F05 model using upsampling and the F10.4 model using downsampling are shown in Fig. 4.1. The F05 model

achieved an AUROC of 0.85 and the F10.4 model an AUROC of 0.93. Model calibration was better for the F05 model than for the F10.4 model.

For the calculation of the variable importance for each model, the *caret* package (including the *randomForest* package) was used. The 30 most important variables for the F05 model and the F10.4 model are illustrated in the Appendix in Fig. A.2 and Fig. A.3, respectively.

4.2 THE DELIRIUM RISK STRATIFICATION ALGORITHM

Finally, both random forest models were integrated into the delirium risk stratification algorithm: the F05 model predicting delirium due to known physiological conditions, and the F10.4 model predicting an alcohol withdrawal state with delirium.

Because a history of delirium is highly associated with current delirium risk (Watt et al., 2018), a rule-based logic was included in the algorithm in addition to the machine learning models: All patients with F05-coded delirium in a previous hospital stay are assigned the highest possible risk score.

4.2.1 *Setting the Thresholds for Stratification*

Before deployment, the thresholds for the risk groups had to be defined. To provide a high clinical utility, existing resources for delirium prevention and the prevalence of delirium were considered for each department. Thus, the thresholds for risk stratification were determined in a discussion with the expert group (see 3.2).

Healthcare professionals agreed on presenting the top 5% of highest rated patients as *very high risk*, followed by the next 10% as *high risk*, and the remaining 85% as *low risk* of delirium. Risk probabilities were predicted with both random forest models on a sub-data set of the participating clinical departments and ranked from the lowest to the highest probability. Cut-offs for the risk groups were set at the 85^{th} and 95^{th} percentile for each model.

4.2.2 *Prospective Risk Stratification*

The prediction by the algorithm is performed as follows: Delirium risk is predicted using the F05 and the F10.4 model. Based on the corresponding model thresholds, a patient is stratified to one of the three risk groups; the higher risk group of both models is used for presentation. In addition, EHR data of the patient are checked for

delirium coded in the past, and patients with history of delirium are stratified to the highest risk class, overruling any prediction by the machine learning models.

In order to initiate a prediction, an HL7 message is automatically sent from openMEDOCS to a communication server for every new admission or transfer to the hospital department. The communication server is monitored by a KAGes internal Linux server with R installation, and the Linux server retrieves all patient data required for modelling from openMEDOCS using http-requests. Risk prediction is then performed on the Linux server using an R environment, and prediction results are sent back to the communication server for visualization in openMEDOCS as well as to a database for documentation. Time stamps, risk probabilities, risk groups and feature values used for modelling are stored in the database every time a prediction is computed.

Risk stratification is repeated in the evening of the admission day in order to include the most recent data from laboratory and nursing assessments.

4.2.3 *Updates of the Algorithm*

After some weeks of prospective prediction during the pilot study, a first update of the algorithm was carried out.

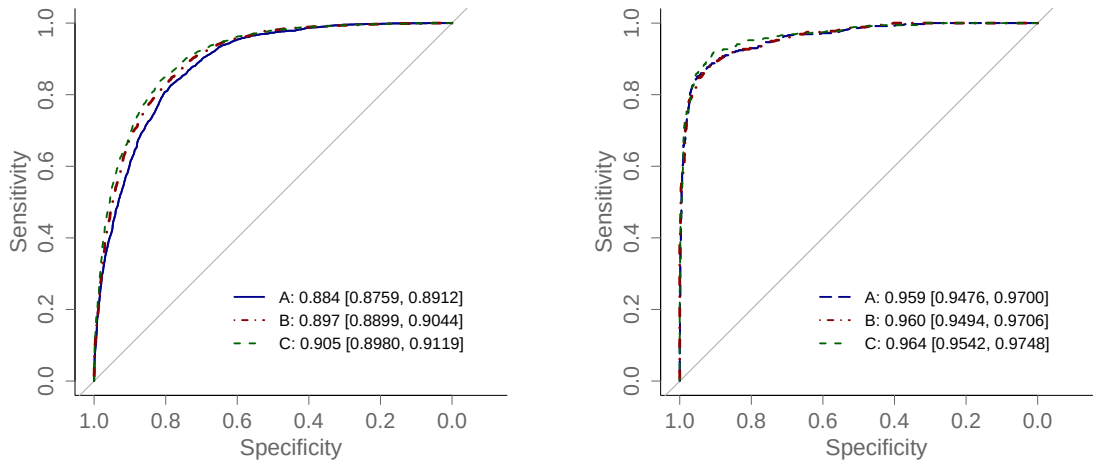
The determination of laboratory parameters is a standard procedure at the beginning of a hospitalisation. However, it is possible that values are available after a few hours. Comparably, an initial nursing assessment is recorded within the first 48 hours. Hence, some information may not be available for prediction at admission.

It was observed that prediction accuracy was limited when specific data were missing for the respective patient. For patients with many missing predictors, e.g. no laboratory values from the last year or nursing assessment data from the last three years, very high risk probabilities were calculated by the random forest models. This was reported as incorrect by healthcare professionals in many cases. Thus, one more rule was added to the algorithm: Patients with less than two predictors available from nursing assessment or laboratory data were assigned a probability of zero.

The first update was supposed to be a quick correction for the problem of missing predictors. A second update of the algorithm in July 2019 addressed the problem in more detail. Prediction models for three different points of time of prediction were trained: Version A for prediction at admission, Version B for 8 pm on the day of admission, and Version C for 8 pm the day after. Each training set included EHR data only up to the defined point of time of prediction in order to simulate the availability of data in clinical routine.

In a third update, drugs prescribed in previous hospital stays were included as a new subgroup of features. As mentioned in Section 4.1.2, information on drugs from admissions in KAGes before 2020 is available from discharge summaries only. Therefore, an information extraction method was used, which (a) extracts drug names from discharge summaries, (b) selects relevant drugs based on a dictionary, (c) assigns drugs to their active ingredients, and (d) assigns these active ingredients to the hierarchical Anatomical Therapeutic Chemical (ATC) classification. Finally, predictors based on the ATC classes were included in the random forest models.

The updates resulted in six different random forest models, three F05 models and three 10.4 models, with excellent and outstanding discrimination on test data sets (see Fig. 4.2). For the latest prediction, Version C, models predicting the outcome F05 and F10.4 achieved an AUROC above 0.90 and 0.96, respectively.



(a) F05 models trained on 29,479 hospitalisations.

(b) F10.4 models trained on 9,134 hospitalisations

Figure 4.2: ROC plots of updated random forest models for three different points of time A, B and C.

In accordance to the changes in model development, time of prospective risk stratification was also adapted. After two predictions on the first day, another recalculation was performed on the second day of the hospital stay, resulting in a minimum of three risk predictions for each admission.

4.3 VISUALIZATION

The predicted delirium risk is displayed to healthcare professionals using two presentation methods. First, a symbol indicating the stratified risk group is displayed in the

user interface of the HIS. Second, patient-specific features relevant for prediction are presented in a web application accessible only within the KAGes network.

4.3.1 *Risk Stratification in the HIS*

In the user interface of the HIS, a column named *Prognose* (German for prediction) is introduced for every participating department (see Fig. 4.3). A red icon symbolizes patients at *very high risk* and a yellow icon those at *high risk*. In order to avoid the presentation of too many symbols in the column of the interface, no symbol is shown for *low risk* patients.

During the pilot study, a question mark icon was added for patients with missing risk prediction. Expert group members agreed that patients without prediction, e.g. due to technical problems, should not be confused to be at *low risk*.

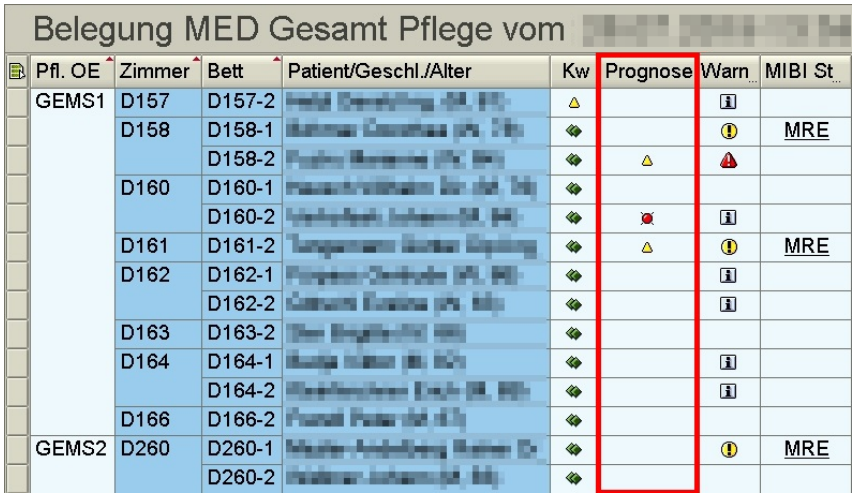


Figure 4.3: Presentation of the delirium risk stratification in the user interface of the KAGes HIS openMEDOCS (Jauk et al., 2020). Symbols in the column "Prognose" indicate a patient's risk group.

4.3.2 *Transparent Visualization of Data Used for Risk Prediction*

Clicking the icon in the HIS (or the empty field for *low risk* patients) opens a web application that provides details on the prediction. The application, illustrated in Fig. 4.4, was developed using the R package *shiny*.

One goal of the delirium risk stratification tool is to open the black-box of machine learning, a problem discussed in Section 2.2.3. Thus, the aim of the web application is to reveal prediction details in order to support clinical reasoning; it displays a patient's EHR data that were used for prediction.

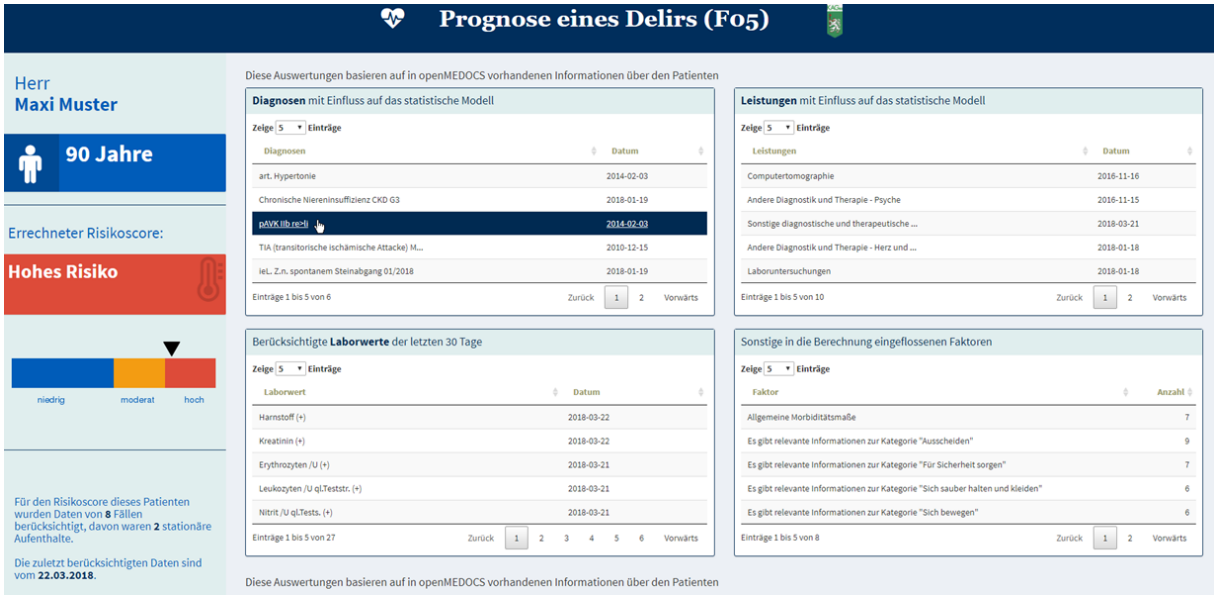


Figure 4.4: Web application visualizing a patient's EHR data used for the prediction of delirium.

The left panel of the web application shows demographic data of the patient as well as the predicted risk of delirium. The risk presentation was changed after a feedback from the expert group during the pilot study, described in Section 6.3.2. The latest presentation includes the colour of a patient's risk group and a bar chart with an arrow indicating the location of the patient on the risk dimension.

In the main panel, EHR data are clustered into four boxes:

- Previous ICD-10 codes with corresponding short texts from clinical notes
- Abnormal laboratory results from the last 30 days
- Procedures from previous hospital stays
- Complementary information, including administrative data and the latest nursing assessment

After the update in July 2019, one box with previous drug prescription was added to the web application (see Fig. 4.5).

Within each box, predictors are ranked by their importance for prediction. The ranking is done based on (1) evidence-based risk factors known from literature and (2) the highest impact on the machine learning-based prediction using variable importance methods.

As EHR data of KAGes hospitals are available across the whole network, information from other hospitals are also visualized in the web application. With a click on the information in the boxes, a pop-up window presents all available values from the patient history, e.g. date and location of all assignments of the ICD-10 code.

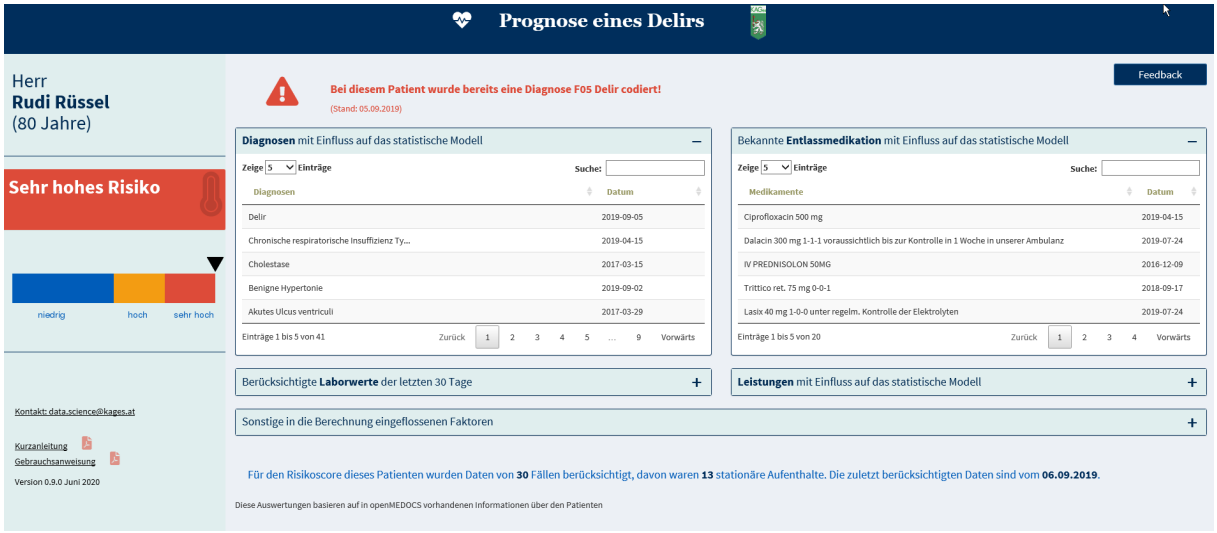


Figure 4.5: Updated version of the web application including previous drug prescription and a feedback button on the top right.

4.3.2.1 *Feedback Button*

To increase the usability of the tool, a feedback button was included on the top right in the web application in July 2019 (see Fig. 4.5). Feedback for individual predictions can be sent directly by the users, as illustrated in Fig. 4.6. Any feedback from users submitted through the feedback button is stored in the database for later analyses.

Figure 4.6: Feedback form included in the user interface of the delirium risk stratification tool. Users can communicate their subjective risk estimation of delirium for a specific patient.

5 PERFORMANCE OF THE ALGORITHM IN A PROSPECTIVE SETTING

This chapter presents the results of the first part of this thesis. It is based on the publication by Jauk et al. (2020).

5.1 INTRODUCTION

Hardly any machine learning-based prediction models have been implemented and publications on the predictive performance of such models in clinical settings are rare. Lee et al. (2020) conducted a systematic review of models based on EHR data and implemented in clinical practice. Although the authors had expected to find a larger number of machine learning-based risk prediction models, only three out of 44 included studies used machine learning methods for modelling.

In the first of these three studies, Hao et al. (2015) developed a random survival forest model predicting 30-day hospital readmission. Their model was successfully integrated into the HIS and achieved an acceptable discriminative performance in a prospective cohort with an AUROC of 0.72.

In the second study, Shimabukuro et al. (2017) validated the performance of a machine learning-based algorithm predicting severe sepsis. The use of the algorithm, which achieved an AUROC of 0.83 on the test data (Calvert et al., 2016), significantly reduced the average length of stay (LOS) and the in-hospital mortality in a randomised controlled trial.

In the third study, Giannini et al. (2019) developed the EWS 2.0 system to predict severe sepsis or septic shock. EWS 2.0, which uses a random forest classifier, achieved an AUROC of 0.88 in a test data set, but its impact on clinical outcomes was limited.

The studies of Shimabukuro et al. (2017) and Giannini et al. (2019) focused not exclusively on the discriminative performance of machine learning algorithms but also evaluated their clinical impact.

One way to validate the clinical utility of an algorithm is comparing its performance with the performance by experts (outlined in Section 2.3.1). Different methods can be used for comparing the performance of humans and algorithms. Magrabi et al. (2019) recommend a comparison of healthcare professionals' decision making with

and without the use of a system, while van Meenen et al. (2014) suggest a comparison between the discriminative performance of a model and physicians' discriminative ability. Most comparisons of algorithms with clinical experts have been conducted with medical imaging use cases in oncology and radiology (Nagendran et al., 2020; Shen et al., 2019).

Brennan et al. (2019) compared the judgment of clinicians with the MySurgeryRisk algorithm predicting several postoperative complications. The assessment was conducted prior to the actual launch of the system; 150 cases of patients were simulated and their risks were estimated by physicians. For five out of six predicted outcomes the algorithm achieved a higher AUROC than physicians in their initial risk assessment. After their interaction with the MySurgeryRisk algorithm, the performance of physicians improved for the risk assessment of acute kidney injury and for ICU admission.

5.1.1 *Aim*

The aim of the first part of this thesis was to evaluate the performance of the delirium risk stratification algorithm in a clinical setting.

The prospective evaluation included (a) an analysis of the predictive performance of the algorithm during during its use in clinical routine, and (b) the validation of its accuracy by comparing the algorithm's outcome with expert ratings.

The comparison was not conducted to demonstrate that machine learning-based risk stratification outperforms healthcare professionals, but rather to validate the results of the algorithm against the gold standard of clinical experts. One goal of the comparison was to detect potential strengths and weaknesses of the algorithm.

5.2 METHOD

In 2018, the delirium risk stratification tool was implemented in a clinical setting for the first time. Two surgical and six internal medicine departments from the KAGes hospital LKH Graz II participated in the pilot study. The timeline of the pilot study design is illustrated in orange in Fig. 3.1 (Section 3.4).

After several training sessions for healthcare professionals, the tool was deployed at the pilot hospital. Since April 2018 delirium risk predictions have been visible in the user interface of the HIS, as described in Section 4.3.

Different methods were used in order to answer the two research questions presented in Section 5.1.1. First, prospective predictions of the algorithm were evaluated over

seven months using measures of discrimination and measures of calibration. Second, prospective comparisons between the algorithm and clinical experts were performed twice in order to validate the accuracy of the algorithm.

5.2.1 *Evaluation of Prospective Predictions*

During the pilot study, delirium risk was predicted for every patient at time of admission or transfer to the participating surgical and internal medicine wards. The risk was recalculated on the evening of the same day. Technical details of the prospective prediction are presented in Section 4.2.2. The specification of the threshold for the risk groups is presented in Section 4.2.1, and measures of performance are described in Section 3.1.

Prospective predictions of the algorithm were evaluated from the 1^{st} of June, 2018, until the 31^{st} of December, 2018. Predictions were excluded from the analysis for patients younger than 18 years.

As illustrated in Section 3.3, the identification of delirium patients in the EHR system was carried out using two methods. First, all those patients were included in the delirium group who had a record of a delirium-related ICD-10 diagnosis (F05, including all subcategories, or F10.4) during the hospital stay and up to 14 days after discharge. Second, discharge summaries from the hospital stay and up to 14 days after discharge were screened for the list of defined words related to delirium. All notes with a positive screening result were manually checked, and patients with clear evidence of delirium were added to the delirium group.

The analysis of the prospective predictions was performed at the level of hospitalisation and, therefore, patients might have been included multiple times due to separate stays. For some hospitalisations more than two risk predictions were available due to transfers between the wards; for them, only the latest prediction within the first 24 hours of the hospital stay was used.

The risk groups *high risk* and *very high risk* were combined for analysis, based on the assumption that they included the top 15% of patients with highest delirium risk. The threshold separating the *low risk* group from the combined *high risk* and *very high risk* group was used for the calculation of discrimination measures: Sensitivity, specificity, false positive rate, false negative rate, positive predictive value and negative predictive value were calculated.

After a detailed analysis of the prospective predictions, the performance of the algorithm on prospective data was compared with the performance of the models F05 and F10.4 on the test data sets (described in Section 4.1.2.2). As measure of discrimination,

ROC curves with a 95%-CI were used. In addition, a calibration plot with a 95%-CI was computed using percentiles of the random forest probabilities. Because the evaluation focused rather on the stratified risk groups than on risk probabilities, measures for calibration were also calculated for the three corresponding risk groups.

5.2.2 *Comparison with Expert Ratings*

In order to compare expert ratings with the algorithm's risk prediction, a protocol was developed for the clinical risk assessment. The assessment was conducted by experienced ward nurses within the first 24 hours of a patient's stay in the respective ward.

The protocol included one item with a five point Likert-type response scale rating the risk of delirium from *very low* to *very high*. Furthermore, all five items of the CAM (Inouye et al., 1990) were included in the protocol. CAM was used to identify patients with current delirium, in addition to standard criteria used for delirium diagnosis in the pilot hospital. The German version of the protocol is included in Appendix A.4.

During the pilot study, the comparison was carried out twice: a first time before the delirium risk stratification tool was visible in the HIS, and a second time after five months of deployment.

In February 2018, one ward nurse completed the protocol for all patients admitted to her general internal medicine ward over a period of 14 days (*Comparison I*). As the tool had not been implemented by then, this comparison was blinded. It provided a first quality assessment of the algorithm's accuracy for a total of 33 patients.

In September 2018, the comparison was repeated for another 14 days (*Comparison II*). This time, the comparison was non-blinded as the tool had been available since spring of the same year. Comparison II was performed by the ward nurse from Comparison I, and by a second ward nurse from a gastroenterolgy ward in the pilot hospital. They provided assessments for 86 general internal medicine and gastroenterology patients that were admitted to one of the two wards. The results of all protocols were analysed using descriptive statistics.

During both assessments, the delirium risk stratification algorithm prospectively predicted the risk of delirium for all patients, as described in Section 4.2.2. For each patient, the latest risk probability and risk group within the first 24 hours of the department stay were retrieved for analysis.

The expert ratings in five categories were compared to the random forest probabilities using boxplots. Relationships between the expert ratings and the algorithm were

evaluated with Spearman's rank correlation coefficient (r), with statistical significance defined at an alpha level of 0.05.

In addition, patients with contradictory risk estimations from experts and algorithm were analysed qualitatively in order to gain insights into the strengths and weaknesses of the algorithm. The amount of available data in the HIS as well as delirium relevant variables were analysed in detail.

5.3 RESULTS

5.3.1 *Evaluation of Prospective Predictions*

During the seven-month evaluation period, delirium risk was prospectively predicted for 5,647 admissions of 4,765 patients. Four patients were younger than 18 years and therefore excluded from the analysis.

For ten admissions, the highest risk group should have been assigned automatically by the algorithm because of a record of delirium in the patient history prior to hospitalisation. However, due to changes in the software code, an error occurred between middle of July and beginning of August, and this part of the algorithm did not execute. All ten admissions were excluded from the analysis.

103 admissions had a probability of zero, which was due to the first update of the algorithm described in Section 4.2.3. For these admissions, no nursing assessment and laboratory data were available from the patient history at time of admission. Thus, their risk was automatically set to zero in order to avoid false positive cases without an underlying clinical explanation. These 103 admissions were also excluded from the analysis.

The final data set included 5,530 admissions from 4,663 patients. The median age of all patients was 71, with an IQR ranging from 57 to 80. Further descriptive statistics for the admissions are presented in Table 5.1. 104 admissions (1.9%) showed a record of delirium in their patient history, and 28 (26.9%) of them developed delirium again during their current admission. Dementia had been coded previously for 4.4% of the admissions, and alcohol abuse for 4.3% of them.

Results of the prospective prediction are presented in a confusion matrix in Table 5.2. Out of all 5,530 admissions, 81.4% were predicted a *low risk* by the algorithm and 18.6% a *high risk* or *very high risk*.

For 81 admissions, an occurrence of delirium was recorded in the EHR system as ICD-10 code or a mention in discharge summaries. Thus, the incidence in the prospective prediction cohort was 1.5%. Of the cases with delirium, 14 were coded

Table 5.1: Descriptive statistics for the admissions during prospective prediction (n = 5,530).

		n	%
Sex	male	2,924	52.9
	female	2,606	47.1
Previously coded diagnoses	Delirium	104	1.9
	Dementia	245	4.4
	Parkinson's disease	87	1.6
	Depression	574	10.4
	Alcohol abuse	236	4.3
	Substance abuse (excl. alcohol, nicotine)	98	1.8

Note: Values are presented as absolute frequencies (column percentages).

as alcohol withdrawal with delirium (ICD-10 code F10.4). The remaining cases were either coded as F05 or type of delirium was unspecified.

5.3.1.1 *Measures of Discrimination*

As illustrated in Table 5.2, 60 out of 81 delirium cases were stratified to the *high risk* or *very high risk* group by the algorithm. This resulted in a sensitivity of 74.1% and a false negative rate of 25.9%, with 21 of 81 delirium cases undetected by the algorithm. Out of 5,449 admissions without indication of delirium, 4,479 were stratified to the *low risk* group by the algorithm. Thus, the specificity for the prospective data was 82.2%, and the false positive rate 17.8%. Positive predictive value was 0.058 and negative predictive value was 0.995.

Table 5.2: Confusion matrix for admissions with prospective prediction of delirium during the pilot study (Jauk et al., 2020).

		Predicted					
		No delirium *(Low)*		**Delirium** *(High, Very High)*		**Total**	
		n	%	n	%	n	%
Outcome	**No delirium**	4,479	82.2	970	17.8	5,449	100.0
	Delirium	21	25.9	60	74.1	81	100.0
	Total	4,500	81.4	1,030	18.6	5,530	100.0

Note: Values are presented as absolute frequencies and row percentages.

In Fig. 5.1, the ROC curve of the algorithm on the prospective data in red is compared with the ROC curves of the random forest models F05 and F10.4 on test data in blue and green, respectively. The discriminative performance of the algorithm on the prospective data (AUROC = 0.855, 95%-CI: 0.8146-0.8956) was as good as for the F05 model (AUROC = 0.854, 95%-CI: 0.8431-0.8642), but lower than for the F10.4 model (AUROC = 0.929, 95%-CI: 0.9122-0.9459).

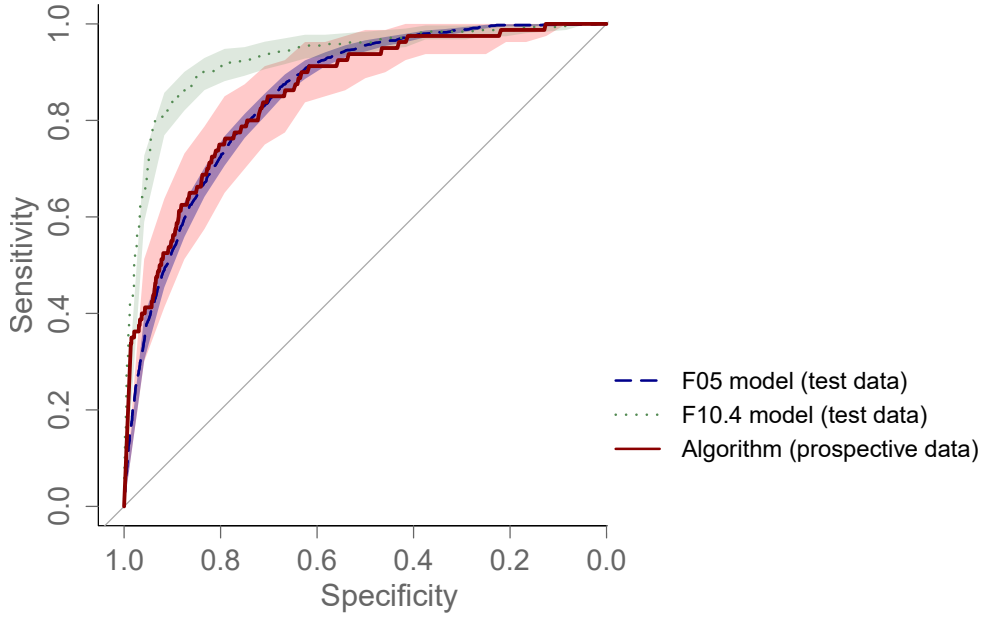


Figure 5.1: Discriminative performance of the algorithm during prospective prediction (Jauk et al., 2020). The ROC curve of the algorithm on prospective data is compared with ROC curves of the models on the test data sets. Curves include 95%-CIs.

5.3.1.2 *Measures of Calibration*

The calibration plot including 95%-CIs is shown in Fig. 5.2. While the calibration of both random forest models was good on the test data, the algorithm achieved only a poor calibration on the prospective data. The average predicted risk of the algorithm was higher than the overall event rate during the prospective prediction for all percentiles of predicted probabilities.

Table 5.3 illustrates the frequencies of delirium within each of the three risk groups for the prospective data and the test data sets. As an example, the thresholds of the internal medicine department were used with cut-offs at 0.576 and 0.714 for the three risk groups. The frequency of patients with delirium in the prospective data deviated strongly from the frequency in the test data sets: While in the prospective data the total frequency of delirium was 1.5%, in the test data of the models F05 and F10.4 it

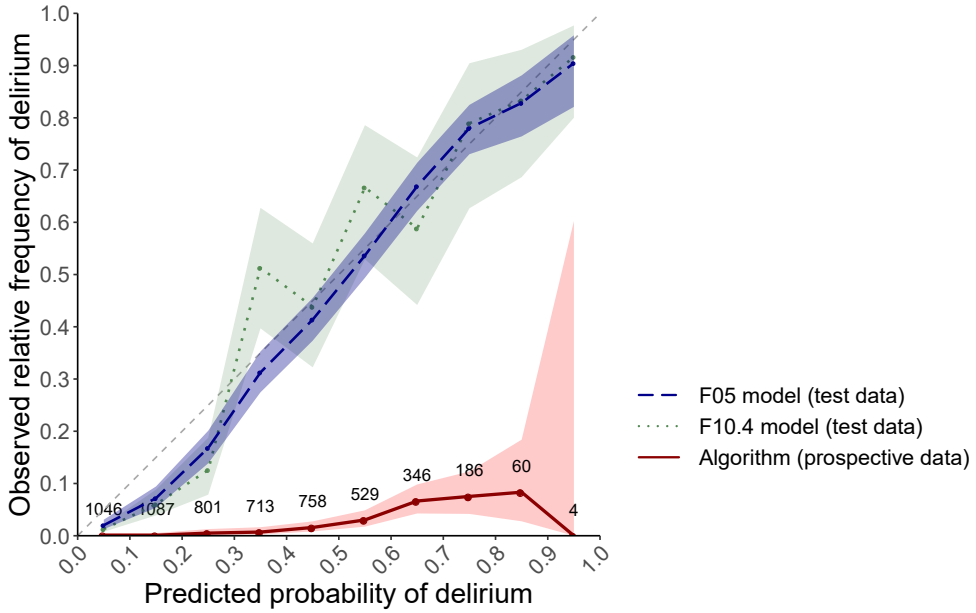


Figure 5.2: Calibration plots of the algorithm during prospective prediction (in red) compared to the two random forest models on test data sets (adapted from Jauk et al. (2020)). Calibration curves include 95%-CIs.

had been 32.6% and 12.6%, respectively. The high frequency of delirium was enforced during the development of the models in order to receive a better outcome balancing.

In the test data, the frequency of delirium cases was in concordance with the percentiles of each risk group, whereas for the prospective data the frequencies were much lower. However, frequency was highest in the *very high risk* group with 9.9% having a record of delirium, and lowest in the *low risk* group with 0.5%.

5.3.2 *Comparison with Expert Ratings*

Descriptive statistics of all patients included in Comparison I and Comparison II are presented in Table 5.4. For the blinded Comparison I, a single ward nurse observed 33 patients within 24 hours after their admission to her ward. One patient had a positive CAM assessment, but no patient was coded as delirium.

Risk ratings and predicted probabilities for all patients from Comparison I are illustrated in Fig. 5.3a. Expert ratings correlated highly positive with the algorithm's risk prediction ($r = 0.81$, $p < 0.001$). For all patients classified as *very low risk* by the nurse (n = 11), the estimation of the algorithm was low as well.

For Comparison II, two ward nurses observed 86 patients. Two patients of Comparison II were coded as delirium, and eleven patients had a positive CAM assessment.

Table 5.3: Frequencies of delirium in the three risk groups for test data F05, test data F10.4 as well as for the prospective data (adapted from Jauk et al. (2020)). Thresholds for internal medicine departments were used.

		Predicted risk			
		Low	**High**	**Very High**	**Total**
Test data F05	N	3,854	593	528	4,975
	With outcome (n)	806	382	436	1,624
	With outcome (%)	20.9	64.4	82.4	32.6
Test data F10.4	N	2,276	69	123	2,468
	With outcome (n)	164	44	104	312
	With outcome (%)	7.2	63.8	84.6	12.6
Prospective data	N	4,500	585	445	5,530
	With outcome (n)	21	16	44	81
	With outcome (%)	0.5	2.7	9.9	1.5

Eight patients (9.3%) had been coded as delirium in a previous hospital stay, but none of them developed a delirium during the current hospital stay.

Ratings and probabilities for Comparison II are shown in Fig. 5.3b. Expert ratings correlated positively with the algorithm's risk probabilities ($r = 0.62$, $p < 0.001$). Again, all patients identified as *very low risk* patients by the nurses (n = 24) were accordantly predicted a low risk probability by the algorithm.

5.3.2.1 *Patients at High Risk of Delirium*

The only patient from Comparison I who had a positive CAM assessment was rated as *very high risk* by the nurse and *high risk* by the algorithm (Fig. 5.3a). The patient was first transferred to a geriatric department of KAGes from a nursing home due to aggressive behaviour and disorientation. After treatment for heart failure and cardiac decompensation in the internal medicine department, he was re-transferred to the geriatric department. The patient was coded as having late onset Alzheimer's disease (ICD-10 code G30.1) and showed aggressive and disorientated behaviour during the whole hospital stay.

Eleven patients (12.8%) of Comparison II had a positive CAM assessment (Fig. 5.3b). The nurse rated all of them as *high risk* or *very high risk*. The algorithm stratified four patients to the *low risk* group, two to the *high risk* group and five to the *very high risk* group.

Table 5.4: Sex, age and previously coded diagnoses stored in the HIS for admissions included in Comparison I and II (adapted from Jauk et al. (2020)).

		Comparison I		**Comparison II**	
N		33		86	
Age, years[a]		71	(63-80)	77	(64-84)
		n	**%**	***n***	**%**
Sex	male	16	48.5	44	51.2
	female	17	51.5	42	48.8
Previously coded diagnoses	Delirium	0	0.0	8	9.3
	Dementia	2	6.1	10	12.8
	Parkinson's disease	0	0.0	3	3.5
	Depression	7	21.2	16	18.6
	Alcohol abuse	3	9.1	8	9.3
	Substance abuse[b]	1	3.0	1	1.2

Note: Values are presented as absolute frequencies (column percentages). [a] Median (Q1-Q3); [b] excluding alcohol, nicotine

Two patients from Comparison II were coded as delirium in the current hospital stay. The first delirium patient had a positive CAM assessment and was rated a *very high risk* by the nurse, but a *low risk* by the algorithm. The risk probability of the random forest model (p = 0.53) was close to the boundary with the *high risk* category. This patient had no hospitalisation within the last three years and thus data for prediction were limited.

The second delirium patient was stratified to the *very high risk* group by the algorithm. The nurse commented that she did not feel confident to rate the delirium risk because of the sedation of this patient. Therefore, she rated the patient with *moderate risk* of delirium. The CAM assessment for the patient was negative.

Six patients of Comparison II had a record of delirium in their patient history. As defined during the algorithm development, all of them were stratified to the *very high risk* group. The nurses rated them either as *very high risk* or *high risk* patients. Although none was coded as delirium during the hospital stay, four had a positive CAM assessment.

5.3.2.2 *Contradictions Between Experts and Algorithm*

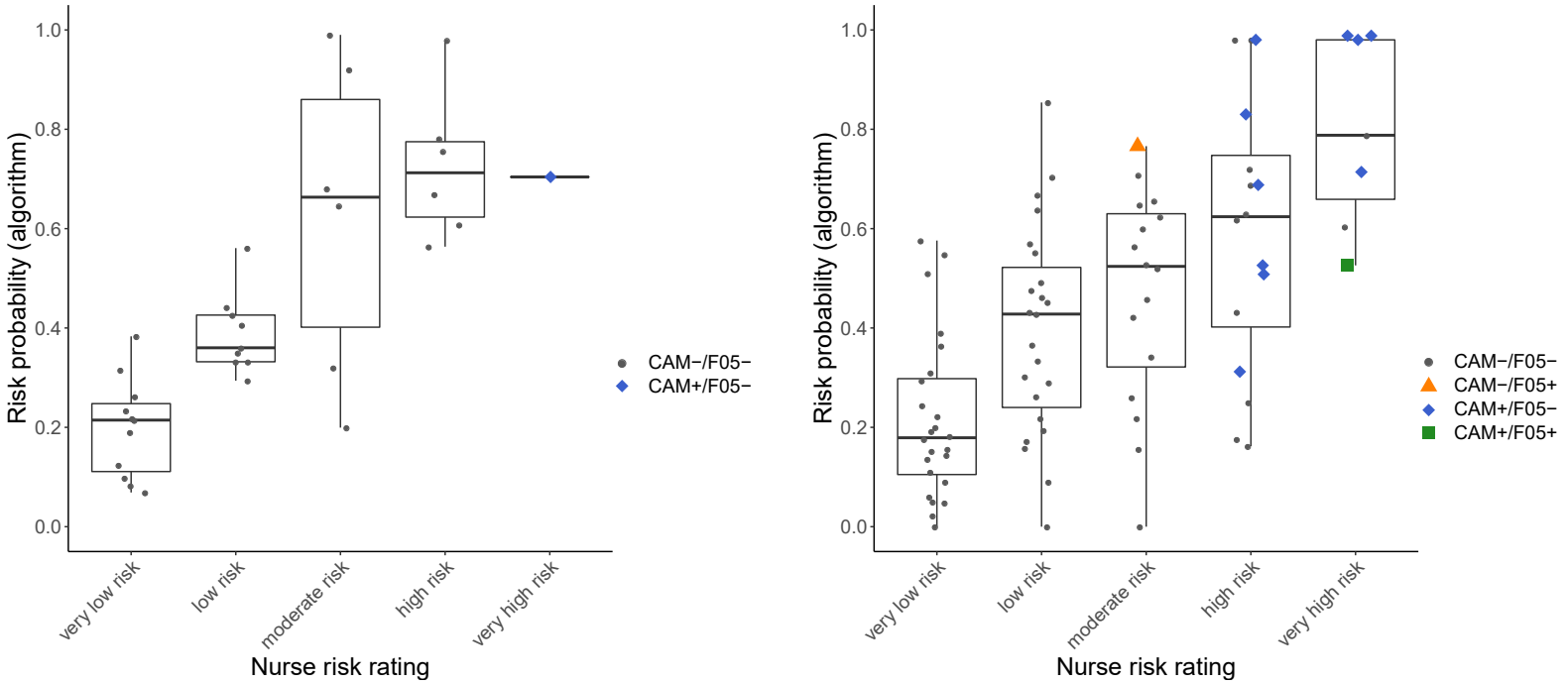
Expert ratings for four patients were lower (*low risk*) than the algorithm's ones (*high risk* or *very high risk*). The patient stratified to the *very high risk* group by the algorithm

had various delirium-related diagnoses in the past, including dementia, as well as observations of disorientation recorded for the current hospitalisation.

Seven patients were considered *high risk* by the nurse, but *low risk* by the algorithm. Three of them had a positive CAM assessment, but none was coded as delirium during the hospital stay. Two patients showed risk probabilities close to the boundary of the *high risk* group; for the third patient little data were available in the HIS.

Three others were relatively young (62 – 64 years), also with little data available in the HIS, and one of them was coded alcohol withdrawal with delirium (ICD-10 code F10.4). One patient had been transferred from the critical care department to the internal medicine department and received mechanical ventilation. The nurse reported that she was unsure about her rating due to the patient's long stay in critical care, and did not assess delirium with CAM.

In summary, the qualitative analysis showed two main reasons for contrary risk estimations: First, in some cases with little data stored in the EHR, the algorithm's predictions were low whereas the nurses' estimations were high. Second, for some other cases the nurses commented that they were unsure about the delirium risk. One reported reason for this uncertainty was a communication barrier due to sedation or language. The algorithm predicted high risk probabilities for most of these cases.



(a) Blinded Comparison I (n = 33).

(b) Non-blinded Comparison II (n = 86).

Figure 5.3: Comparisons of nurses' risk ratings and risk probabilities of the algorithm for general internal medicine and gastroenterology patients (adapted from Jauk et al. (2020)). (a) Comparison I was conducted for one ward in February before the tool was available in the HIS. (b) Comparison II was conducted in September for two wards. Patients with positive CAM assessment and/or ICD-10 coded delirium (F05) are highlighted.

5.4 LIMITATIONS

Several limitations to the first part of the thesis need to be acknowledged. The biggest limitation of the pilot study was the low incidence of delirium. For more than 5,530 hospitalisations evaluated during seven months, delirium could be identified in the EHR data in 1.5 % of the cases. As illustrated in section 2.4, published studies and guidelines report an incidence of delirium up to 30%. Thus, the incidence observed in the prospective evaluation cohort was lower than expected. Because this problem presents a major limitation for this thesis it is further elaborated in Section 8.2.

5.4.1 *Limitations of the Prospective Prediction Results*

The low incidence of delirium in the prospective data led to non-informative calibration plots and limits the interpretation of measures for discrimination. During the prospective performance, only 14 admissions were coded as alcohol withdrawal with delirium (F10.4). This might be one reason why the AUROC of the algorithm in the prospective data was lower than the AUROC of the F10.4 model, which was higher than the discriminative performance of the F05 model on the test data.

Two limitations of the study might have led to an overestimation of the false positive rate of 0.178. First, the clinical setting did not allow controlling for interventions to prevent delirium. If prevention is successful, a delirium occurs less frequent. A self-destroying prophecy, a phenomena known from sociology (Sabetta, 2019), is then observed: The algorithm predicts a high risk of delirium, and delirium is successfully prevented due to interventions in the ward. In the analysis, however, such patients are categorized as false positive.

Second, patients from 104 admissions (1.9%) had a delirium diagnosis in their past and were classified as *very high risk*. Out of them, 28 (26.9%) patients developed delirium again during their current admission. Although the patients of the remaining 76 cases did not develop delirium, they were seen as patients with highest risk per definition (see Section 4.2).

An essential limitation encountered during prospective prediction is related to the data available for prediction. Previous research had shown that clinical decision support based on EHR can be influenced by late data entry (Perry, Hossain, & Taylor, 2018). Late data entry or delayed availability also influenced this pilot study, and certain EHR data used for prediction were missing not at random. When a patient with an existing KAGes-EHR is admitted to the hospital, most data are available at the time of admission, but latest laboratory data and nursing assessment might need

some time for evaluation or data transmission. When a patient is admitted to one of the KAGes hospitals for the first time, most features are not available.

A simulation study on KAGes data demonstrated that for patients with missing laboratory data and nursing assessment a model trained specifically for that scenario discriminates better than a model trained with all features (Jauk et al., 2019). Thus, one way to account for time delays in information availability is the use of specifically trained models depending on the data availability. As a result, the delirium risk stratification algorithm was updated accordingly in July 2019 (see Section 4.2.3).

5.4.2 *Limitations of the Comparison With Expert Ratings*

In order to validate the accuracy of the algorithm, predictions were compared to risk ratings from clinical experts. The first comparison, Comparison I, took place before the implementation of the delirium risk stratification tool and was thus blinded for the expert. This was essential, because the expert rating was not influenced by any risk visualization in the user interface of the HIS. In contrast, for Comparison II the risk ratings of the nurses could have been biased, even though the nurses reported that they intended not to be influenced by the tool.

Only two patients of both comparisons were coded as having delirium, both without previous delirium diagnoses in their patient histories. One of the delirium patients was detected by a nurse, the other one by the algorithm only. Because of this low incidence, a comparison of sensitivity between nurses and algorithm was not informative.

The comparisons were also limited by two aspects regarding the item used for the experts' risk rating. First, both nurses reported that they used the category *moderate risk* whenever they were unsure about the delirium risk. Although this supports the assumption that the tool provides an additional support in delirium management for cases of uncertainty, results for patients rated as *moderate risk* are limited.

Second, while the algorithm stratified patients into three risk groups, the item of the expert rating had five risk categories. Hence, the three risk groups of the algorithm could not be compared to the five risk groups from expert ratings without restrictions. For quantitative analyses in Comparison I and II, probabilities of the algorithm were used instead. The interpretation of these probabilities is limited by the poor calibration of the algorithm for prospective data.

Many studies have investigated verbal and numeric scales for risk estimation and their uncertainty, and there is slight evidence that verbal scales result in more reliable risk estimations than numerical ones (Ancker et al., 2006). Therefore, the use of verbal

risk categories instead of numerical risk percentages seemed appropriate for the item in the protocol.

Finally, the five point Likert-type scale was chosen for two reasons. First, five-point or seven-point scales are preferred over three-point scales in literature (Diefenbach, Weinstein, & O'Reilly, 1993). Second, the use of a three point Likert-type response scale with approximately equal intervals would not have been comparable to the three risk groups of the algorithm representing the upper 5%, 10% and 85% of delirium risk.

It was out of scope of this thesis to analyse the rating scale using methods of item response theory (e.g. a Partial Credit Model), but it should be included in future research.

6

TECHNOLOGY ACCEPTANCE

As much as an algorithm excels in prospective prediction, it is crucial to know how users and domain experts perceive it and how they adopt it in their daily routine. This chapter presents the evaluation of the technology acceptance of the delirium risk stratification tool by healthcare professionals, and is based on the results published by Jauk et al. (2021).

6.1 INTRODUCTION

6.1.1 *The Technology Acceptance Model*

A well-known model for evaluating the acceptance and rejection of technologies is the Technology Acceptance Model (TAM). TAM was originally developed by Davis (Davis, 1989; Davis, Bagozzi, & Warshaw, 1989), and has been widely used and adapted during the last decades. The model is often referred to as a gold standard for explaining IT acceptance (Holden & Karsh, 2010).

TAM is based on the theory of reasoned action by Fishbein & Ajzen (1975) which assumes that an intention for a certain behaviour or human action is influenced by one's attitudes. This intention acts as a predictor for the actual behaviour. Accordingly, TAM assumes that a behavioural intention acts as best determinant for the actual use of a technology.

The behavioural intention is influenced by two factors, *perceived ease of use* and *perceived usefulness* of a system (Davis, 1989). Perceived ease of use is defined as "the degree to which a person believes that using a particular system would be free of effort". Perceived usefulness refers to "the degree to which a person believes that using a particular system would enhance his or her job performance". Perceived ease of use is also linked to intention via an impact on perceived usefulness.

An extension of the model, the Technology Acceptance Model 2 (TAM2), was developed by Venkatesh & Davis (2000). In TAM2, perceived usefulness is further influenced by seven additional factors such as experience, job relevance, voluntariness and output quality. *Output quality* of a system is defined by the authors as the perception of users how well a system performs its tasks.

Validity and robustness of TAM have been shown in several contexts (King & He, 2006). Although TAM was originally developed for the evaluation of electronic mailing systems, it has been widely used as a theory for evaluating healthcare technologies. Several examples for healthcare evaluations using TAM as an underlying construct can be found in a review by Holden & Karsh (2010). The authors point out that the generic form of TAM might not capture all features specific for computerized healthcare applications. Thus, a slight adaptation of the items assessing TAM is recommended when evaluating health IT applications.

6.1.2 *Aim*

The aim of the second part of this thesis was to gain knowledge of the user acceptance, uptake and concerns regarding the delirium risk stratification tool. The evaluation targeted domain experts as well as those physicians and nurses who had been using the tool during the pilot study.

A major objective of the tool is to provide an explanation of the prediction in order to support decision making of healthcare professionals. This is attempted by the visualization in the HIS and in the web application described in Section 4.3.2. Thus, the evaluation of technology acceptance involved not only the output of the delirium risk stratification algorithm, but also the visualization of the results.

6.2 MATERIALS AND METHOD

6.2.1 *Study Design*

For the mixed method study, a convergent parallel design was used (illustrated in Fig. 6.1). TAM (Davis, 1989) constituted the evaluation framework for the quantitative and qualitative assessment.

In addition to the TAM factors *perceived ease of use* and *perceived usefulness*, the factor *output quality* from TAM2 (Venkatesh & Davis, 2000) was used for evaluation. In this study, output quality was defined as the perceived correctness of the predicted delirium risk.

According to TAM, the intention to use a system is influenced by these three factors, and will further lead to the actual behaviour, also defined as *actual system use*. Fig. 6.2 shows the interaction of the TAM and TAM2 factors used in the study.

The timeline of the qualitative and quantitative assessment is illustrated in blue in Fig. 3.1 (Section 3.4). For the qualitative assessment, opinions and comments were

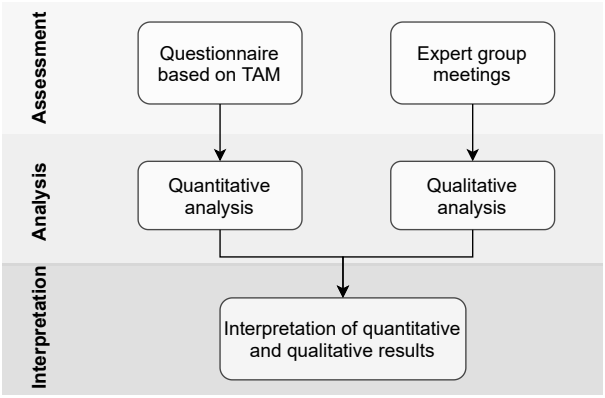


Figure 6.1: Convergent parallel study design for the assessment of technology acceptance using quantitative and qualitative methods (adapted from Jauk et al. (2021)). TAM by Davis (1989) served as a framework for evaluation.

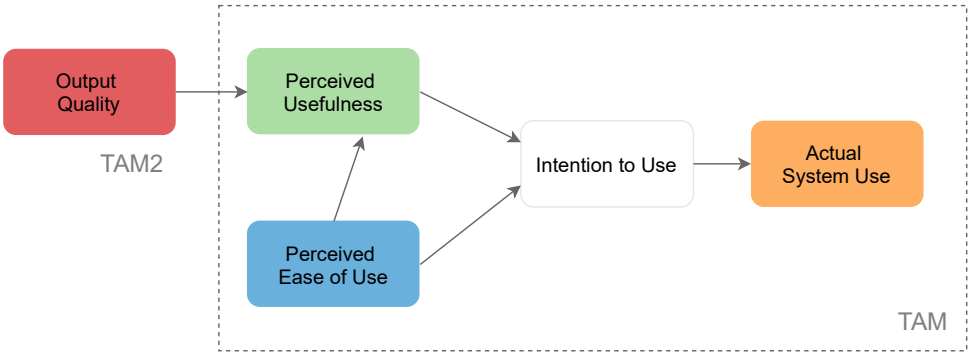


Figure 6.2: Factors of TAM (Davis, 1989) and TAM2 (Venkatesh & Davis, 2000) framed the evaluation of technology acceptance.

collected from physicians and nurses participating in the expert group, described in Section 3.2. The first expert group meeting took place in November 2017, followed by another three meetings until December 2018. Field notes were taken during all meetings and, finally, all collected comments were assigned to the four factors of TAM and TAM2.

The aim of the quantitative assessment was to receive a broad feedback on the delirium risk stratification tool. A technology acceptance questionnaire was used for the assessment. From October until December 2018, printed questionnaires were distributed to healthcare professionals working in five of the eight wards which participated in the pilot study. Physicians and nurses from all levels of experience were encouraged to participate in the assessment in order to receive feedback from different user groups. The participation was anonymous and on a voluntary basis.

Finally, all results from the quantitative and qualitative assessment were interpreted in conjunction, in order to obtain a detailed picture of the acceptance of the tool in clinical routine.

6.2.2 *The Technology Acceptance Questionnaire*

As TAM and TAM2 were developed outside of healthcare, an adaptation to the particular context of an evaluation is recommended (Holden & Karsh, 2010). Thus, several sample items were formulated for each TAM factor based on original examples by Venkatesh & Davis (2000). The items were slightly adjusted to the context of healthcare in general and to delirium prediction in particular.

One goal of the development was to use as few items as possible considering the restricted time resources of healthcare professionals in hospitals. Based on internal discussions, 16 items were selected for the questionnaire. A pilot test on the understandability of all items was performed with two hospital staff members which were not otherwise involved in the study. The pilot test resulted in slight adoptions of the formulation for some items.

The final questionnaire included five items measuring perceived usefulness, six items for perceived ease of use, three items for actual system use and two items for output quality. Responses on 15 items were measured using a five point Likert-type response scale from *strongly agree* to *strongly disagree* and from *very frequently* to *very rarely*. The 16^{th} item assessed the absolute frequency of use per month in numbers. In addition, user comments were assessed in a free-text field at the end of the questionnaire. The original version of the questionnaire in German and a forward translation to English are included in the Appendix (see Fig. A.5 and Fig. A.6, respectively).

6.2.3 *Quantitative Data Analysis*

Two items measuring perceived ease of use had been formulated negatively and had to be recoded for analysis. The item with the original formulation *The application was difficult to use* was recoded and interpreted as *The application was not difficult to use*. The item *The application has increased my workload* was interpreted as *The application has not increased my workload* after recoding.

For each participant, the median of the item responses for each TAM factor was calculated, and then the mean of the medians of all participants was calculated for each factor. A mean or median of 1 indicates a positive answer or strong agreement, a mean or median of 5 a negative answer or low agreement.

Heat maps and descriptive statistics facilitated the analysis of the items of each factor independently.

In an exploratory analysis, the internal consistency of the TAM factors in the questionnaire was assessed using Cronbach's alpha. Cronbach's alpha (Cronbach,

1951) determines the reliability of each factor in describing how closely related the items of the factor are. Acceptable values of alpha are usually above 0.7, but too high values (e.g. above 0.9) can indicate redundancies of items (Tavakol & Dennick, 2011).

First, the mean of all items for each factor was calculated. Second, Cronbach's alpha was calculated for each factor using the R package *ltm* (Rizopoulos, 2018), and 95%-CIs were calculated using boosting methods.

6.3 RESULTS

Between October and December 2018, questionnaires were completed by ten out of 21 physicians (47.6%) and 37 out of 67 nurses (55.2%) from five hospital wards. Descriptive statistics for participants of the quantitative assessment are presented in Table 6.1.

Table 6.1: Descriptive statistics of the participants of the quantitative assessment of technology acceptance (n = 47, Jauk et al. (2021)).

		n	%
Profession	Nurse	37	78.7
	Physician	10	21.3
Sex	Male	14	29.8
	Female	33	70.2
		Median	(Q1-Q3)
Age, years		29	(26-42)

Table 6.2 shows the mean responses for each factor, with ratings from 1 (positive/high) to 5 (negative/low). Users rated the perceived ease of use and perceived usefulness rather positive, the output quality neutral, and the actual system use rather poor.

Table 6.2: Mean responses to the factors measuring technology acceptance of the delirium risk stratification tool. Means and standard deviations were calculated for the median response of all participants (Jauk et al., 2021).

TAM factor	**Items (*n*)**	**Mean**	**SD**
Perceived usefulness	5	2.4	1.00
Perceived ease of use	6	2.3	0.79
Output quality	2	2.8	0.73
Actual system use	2	3.4	1.0

Results on internal consistency of the four TAM factors are reported in Table 6.3. The internal consistency was acceptable for perceived usefulness, perceived ease of use and actual system use, but rather low for output quality.

Table 6.3: Internal consistency using Cronbach's alpha for the TAM factors used in the quantitative assessment.

		Cronbach's alpha		
TAM factor	**Items (n)**	**Alpha**	**Lower CI**	**Upper CI**
Perceived usefulness	5	0.76	0.650	0.837
Perceived ease of use	6	0.75	0.604	0.842
Output quality	2	0.66	0.070	0.876
Actual system use	2	0.85	0.743	0.918

6.3.1 *Perceived Usefulness*

A heat map of the items measuring perceived usefulness is shown in Fig. 6.3. In concordance with the positive mean response illustrated in Table 6.2, results of four out of five items were rather positive.

32 users (68.1%) agreed or strongly agreed that the application provided them with additional information. Seven users (14.9%) disagreed or strongly disagreed that the application is a useful support for delirium prevention, and seven did not believe that the application can be used to detect delirium at an early stage. Opinions on the usefulness for the users' own work were mixed: 17 users (36.2%) agreed or strongly agreed that the application is useful for their work, while 15 users (31.9%) disagreed or strongly disagreed on that.

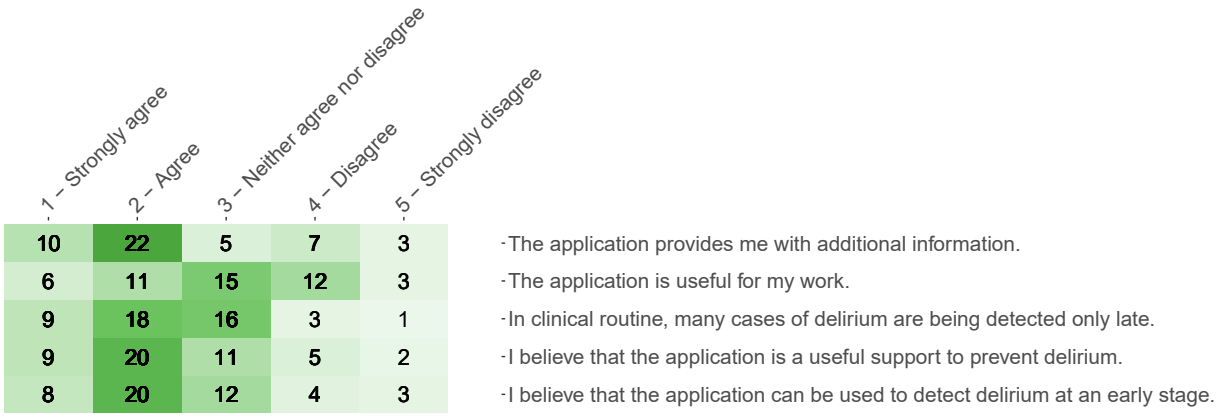


Figure 6.3: Heat map of five items measuring perceived usefulness of the tool (n = 47, adapted from Jauk et al. (2021)).

The consensus of the expert group on perceived usefulness was that the tool offered a great support in early recognition of delirium patients. In their opinion, the risk stratification helps to reduce screening resources.

> "The tool gives good support – I am convinced of its usefulness."
>
> "Due to the delirium prediction tool, we were already able to prevent the sliding into a strong delirium with simple interventions."
>
> "I perceive the tool as a benefit as we are able to reduce the time for delirium screening."

It was observed that for some cases the tool supported the estimation of healthcare professionals strongly, e.g. if patients were under sedation. In addition, the tool was used by healthcare professionals to confirm existing presumptions.

> "It is especially an added value if patients are not responsive during admission."
>
> "The risk prediction helps to corroborate my own estimation when seeing a patient."
>
> "The risk prediction supports us when we are not quite sure about the delirium risk."

The tool also supported healthcare professionals in targeting patients with a delirium diagnosis in a previous stay.

> "Especially patients with a diagnosis of delirium in the past are being targeted earlier now."

6.3.2 *Perceived Ease of Use*

Answers on the questions for perceived ease of use were rated rather positively, as illustrated in Fig. 6.4. Only two users strongly disagreed that the purpose of the tool was clear and understandable, and three users disagreed that the presented information was understandable. Four users reported that the application had increased their workload.

However, 18 users (38.3%) disagreed or strongly disagreed that they successfully integrated the tool into their clinical routine, and 14 users (29.8%) reported that they were not sufficiently prepared to use the tool at time of implementation.

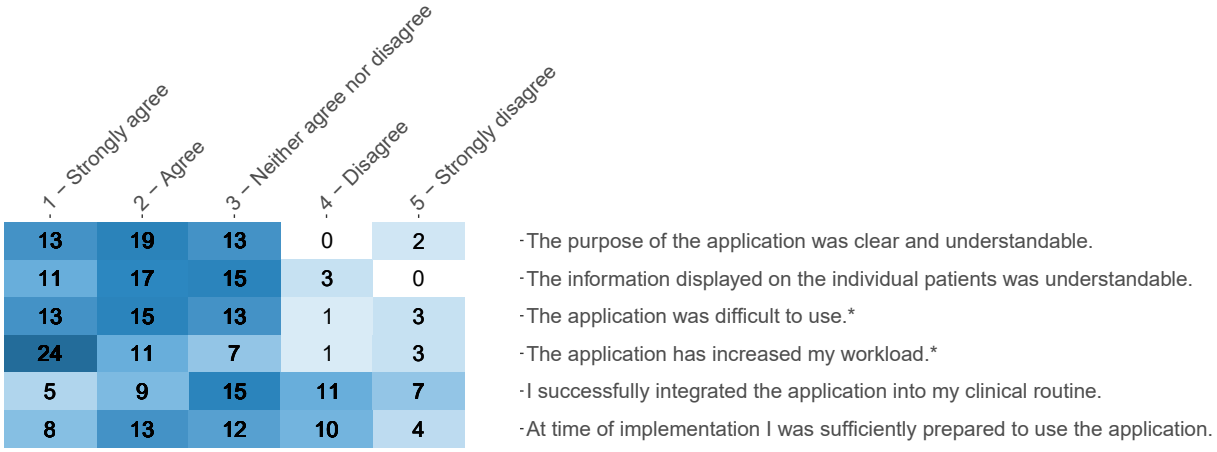


Figure 6.4: Heat map of six items measuring perceived ease of use of the tool (n = 47, adapted from Jauk et al. (2021)). Two items marked with asterisk were recoded for analysis.

In the expert group, the common impression of the perceived ease of use was positive. Clinical experts appreciated that there was no need of additional data entry and that the prediction was available within few seconds in the user interface of the HIS.

As shown in Section 4.3.1, *high risk* patients were presented with a yellow symbol, and *very high risk* patients with a red symbol. This visualisation was appreciated by the expert group.

> "I like the presentation with the traffic light symbols."

The web application sparked much enthusiasm because it provided a comprehensive view of a patient and thus supported healthcare professionals not only in delirium management (see Section 4.3.2). However, during the first month of deployment, the delirium risk had been visualised using percentages. The clinical experts criticized the numbers, because they found them difficult to interpret. As a solution, the percentages were replaced by a bar chart visualizing the risk groups and an arrow indicating the location of a patient on the dimension of delirium risk.

> "The bar representing the range of delirium risk helps us to identify patients at the boundary to another risk group."

6.3.3 *Output Quality*

The heat map for the two items measuring output quality is shown in Fig. 6.5, demonstrating rather neutral results for this factor. For both questions, the middle

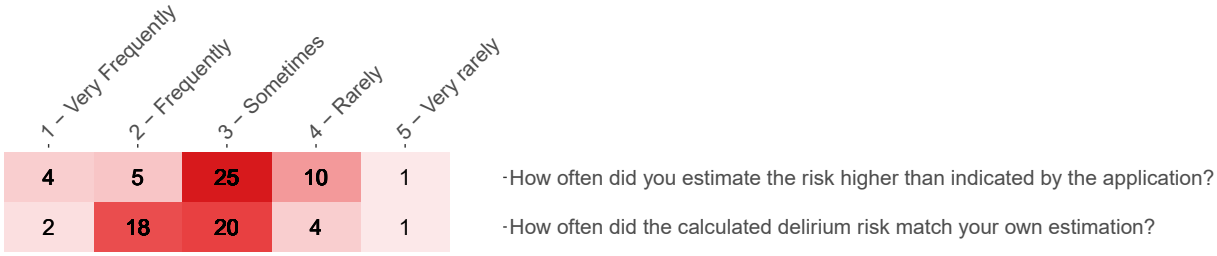


Figure 6.5: Heat map of two items measuring the output quality of the tool (n = 47, adapted from Jauk et al. (2021)).

category was the most chosen answer. 20 users (42.6%) reported that the delirium risk matched their own estimations frequently or very frequently.

Within the expert group, the predictive accuracy of the delirium risk stratification algorithm was perceived as very high.

> "The system has almost 100% accuracy."

> "There are not too many patients in the very high risk group – it seems correct."

6.3.4 *Actual System Use*

Answers on the two items measuring the actual system use were rather negative, as shown in Fig. 6.6. 19 users (40.4%) disagreed or strongly disagreed that they considered the output of the application in their clinical decisions. 13 users (27.7%) agreed or strongly agreed that they had been using the application regularly; the median for use per month was 3 times (min = 0, max = 20).

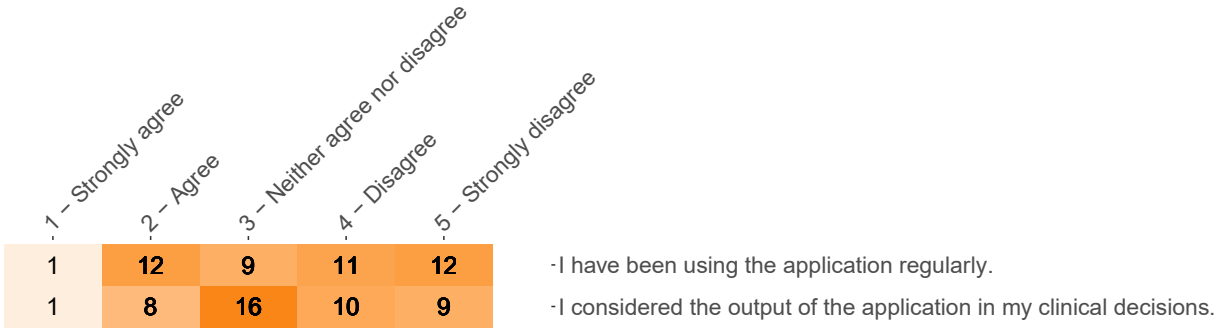


Figure 6.6: Heat map of two items measuring the actual use of the tool (n = 47, adapted from Jauk et al. (2021)).

One senior physician from the expert group raised concerns about the frequency of use among other physicians.

> "I absolutely want to continue with the tool. Now the question is how to bring it closer to the users – many don't know much about it yet."

Finally, there was a broad agreement of the clinical experts that the overall impression of the delirium risk stratification tool was positive and they wanted to continue its use in clinical routine. Also, the expert group recommended the implementation of the tool in other hospital departments and hospital networks.

> "The tool is successful. It should be continued in any case."

6.4 LIMITATIONS

Several aspects limit the second part of this thesis. Besides limitations of the mixed methods design, the study was limited by the use of TAM as a framework for evaluation.

6.4.1 *The Use of TAM in Healthcare*

Although TAM is seen as one of the gold standards in understanding the acceptance or rejection of health IT applications, three main limitations to its use and to the adaptation of its items need to be discussed.

First, the interpretation of the results when using TAM or TAM2 in healthcare is limited. According to a review by Holden & Karsh (2010), perceived usefulness was significantly related to intention to use in all reviewed studies, whereas perceived ease of use was a significant predictor in only half of the studies. Especially if users are experienced with a system or have enough IT support, perceived ease of use might not be an important predictor (Holden & Karsh, 2010, cited by Ammenwerth, 2019). According to Ammenwerth (2019), TAM could also not be able to assess the rather complex reality of a healthcare setting in every evaluation study.

Second, although the review of Holden & Karsh (2010) recommends the adaptation of items to a specific context, key factors of the original TAM model are prone to multidimensionality when using an adapted item pool. For the assessment in this thesis, questionnaire items were adapted to the use case of delirium prediction. Due to the limited sample size, further psychometric analyses such as factor analyses were not feasible. Although the internal consistency measured with Cronbach's alpha was acceptable for the factors used, a more comprehensive analysis on the questionnaire is needed.

Third, as stated by Ammenwerth (2019), TAM "was developed with a focus on technology which can be used voluntarily". The delirium risk stratification tool under evaluation is supposed to provide support in delirium management if needed. However, during the pilot study physicians and nurses were strongly encouraged by the heads of the departments to use the tool, which could have further influenced the results of the assessment.

6.4.2 *Limitations of the Quantitative Analysis*

Besides already discussed limitations regarding the use of TAM, the main limitation of the quantitative assessment was the use of a convenience sample. 48% of the physicians and 55% of the nurses from five hospital wards participated in the assessment, but the absolute number of physicians and nurses was unequally distributed: The questionnaire was completed by 37 nurses and 10 physicians. Even though it is known that the treatment of hospitalised patients with delirium presents a high burden especially for nurses (Schmitt et al., 2019), physicians' experience with the tool should be further assessed.

In addition, some groups of healthcare professionals might be more likely to participate in the evaluation, e.g. those more positive towards an application. This effect, also called participation bias or non-response bias (Michie & Marteau, 1999), certainly affected the results of the assessment.

6.4.3 *Limitations of the Qualitative Analysis*

Two limitations regard the qualitative assessment of this study. First, even though members of the expert group were chosen thoroughly and with respect to different professions, the group consisted of a small and non-representative sample. The hierarchical system, as commonly known from healthcare institutions, could have led to a bias in the collection of information. Unintentionally, comments from heads of medical departments might have received more attention than those from rather young nurses.

Second, no structured interviews were conducted in the expert group, as the aim of the qualitative assessment was to receive a broad feedback without restrictions to specific questions. Comments from clinical experts were documented and all of them could be assigned to the TAM factors chosen for evaluation. Although the selection and assignment of the comments was reviewed by a second researcher, both could be biased.

7 LONG-TERM PERFORMANCE IN A MULTICENTRE SETTING

7.1 INTRODUCTION

With evolving healthcare systems, clinical prediction models can become inaccurate over time (Jenkins et al., 2018). Models are prone to changes in patient populations, treatment possibilities or prevalence of diseases. A machine learning model, which was trained on static data, will certainly not perform well if data needed for prediction are changing.

Systematic changes of clinical documentation in EHR systems present a common threat to model performance. One example are ICD revisions, which include reclassifications of codes, changes in numeric codes or in code descriptions (O'Malley et al., 2005). ICD revisions may bias the analysis of diagnoses stored in EHR systems (Janssen & Kunst, 2004) and may affect a model's performance over time.

In addition, performance can also vary between hospitals. Hospitals often differ in patient populations or in quality of EHR data. As the coding of diagnoses and procedures happens often because of administrative reasons, the precision and also the mindset towards coding can differ even between departments within a hospital. Thus, the external validation of a prediction model in one centre does not guarantee the same results for other centres.

The long-term evaluation of prediction models should consider another effect. If the model deployment and the interventions staged due to a prediction lead to the successful prevention of an outcome, the variance of this outcome and its association with certain predictors can be reduced (Lenert, Matheny, & Walsh, 2019). This can lead to a decreased discriminative performance of the model. Furthermore, prediction models can underlie a calibration drift leading to under- or overestimation of risk and, hence, to misleading results (Jenkins et al., 2018).

One way to account for ongoing changes in EHR data is a scheduled retraining of clinical prediction models. However, the necessity of retraining has to be assessed for each use case in particular. Thus, it is crucial to continue the evaluation of model performance as soon as models are successfully deployed in clinical routine in order to survey the stability of models over time and for different centres.

7.1.1 *Aim*

The third aim of this thesis was to evaluate the long-term performance of the delirium risk stratification algorithm. The main goal was to analyse the stability of its performance when deployed in different KAGes hospitals across Styria over the year 2019.

Two exploratory analyses enriched the evaluation. First, anonymous feedback of healthcare professionals received through the tool was analysed. Second, discharge summaries of identified delirium patients were analysed in order to explore patterns in clinical texts of patients with and without assigned ICD-10 diagnoses for delirium.

7.2 METHOD

7.2.1 *Study Design*

The evaluation of the delirium risk stratification tool in the pilot study resulted in a recommendation by clinical experts to implement it in other hospitals (see Section 6.3.4). Thus, after the pilot phase the tool was deployed in four other KAGes hospitals in Styria.

The timeline of the study design for this part of the thesis is illustrated in Fig. 3.1 (Section 3.4) in red. The time period of deployment is shown for each hospital included in the multicentre evaluation. For the pilot site LKH Graz II, now referred to as Hospital 1, the use of the tool was continued after the pilot study without disruption.

For every hospital, the implementation of the delirium risk stratification tool was carried out based on the experience from the pilot study. Training sessions for physicians and nurses were organized before the actual implementation in each hospital. The training included a short introduction of the tool emphasizing its strengths and weaknesses and a presentation of the main results of the pilot study. The participation was recommended by the heads of the department but remained voluntary.

7.2.2 *Participating Hospitals*

Medical specializations of KAGes hospitals depend on their location in Styria. As the delirium risk stratification algorithm was developed on a cohort of internal medicine and surgical patients, the tool was implemented at hospitals with at least one of the two specializations.

In Hospital 1 (LKH Graz II), all internal medicine departments and surgical departments continued with the tool in 2019. In July 2019, the tool was further implemented at surgical and internal medicine departments of Hospital 2 (LKH Weststeiermark) and Hospital 3 (LKH Rottenmann - Bad Aussee). Hospital 4 (LKH Hochsteiermark) followed with an implementation at the surgical departments in September 2019. Finally, in October 2019 the tool was implemented at the cardiology department in Hospital 5 (LKH-Univ. Klinikum Graz), the university hospital of KAGes.

7.2.3 *Analysis of the Prospective Performance*

Based on the pilot study results, the algorithm was slightly updated in July 2019 (see 4.2.3). The update included a re-training of the random forest models and the deployment of an additional re-calculation on the second evening after admission.

Because of this update, the data of Hospital 1 were split for analysis: The first data set included prospective predictions from January until June 2019 using the first version of the algorithm; the second data set included prospective predictions from August until December 2019 using the updated version of the algorithm. The month of July was excluded in order to control for technical changes and an overlap of the two versions.

Hospital stays of patients with an admission and discharge within the defined time period were included in the analysis. In order to account for possible technical errors at the start of deployment, the first days after the implementation in each hospital were excluded from the analysis. Patients younger than 18 years were excluded from the analysis. Descriptive statistics of the patients were analysed for each hospital. The median LOS was calculated in days, including admissions with hospital stays from one day up to 100 days.

Occurrence of delirium was identified using the two methods illustrated in Section 3.3: ICD-10 codes for delirium recorded in the EHR system and text mining of discharge summaries for indications of delirium.

Discriminative performance and calibration were analysed on a hospital level. Two risk predictions were extracted from the database for each hospitalisation:

(a) the first risk prediction of each hospitalisation, and
(b) the last risk prediction within the first 48 hours of each hospitalisation.

If a patient is not further transferred during an hospital stay, this risk prediction will be visible over the entire stay in the HIS and will not be re-calculated. Thus, for interpretation of the results, the more relevant prediction of both is (b), the last prediction.

AUROC values including 95%-CIs were compared for both risk predictions in order to evaluate the importance of the re-calculation of delirium risk. Further analyses including ROC curves and calibration plots with 95%-CIs were conducted for prediction (b) only.

7.2.4 *Analysis of Feedback from Healthcare Professionals*

In July 2019, a feedback button (described in Section 4.3.2.1) was added in the web application. Users were enabled to send feedback on a patient's prediction, which included their rating of the delirium risk and their comments in a free text field.

The feedback received through the web application was analysed between August and December 2019. Risk groups rated by healthcare professionals were compared with the risk groups predicted by the algorithm. If no risk rating was provided by healthcare professionals, the content of the free-text comments was analysed to identify over- or underestimation of the predicted delirium risk.

7.2.5 *Analysis of Discharge Summaries of Delirium Patients*

Most discharge summaries had already been exported for the evaluation of the long-term performance; in addition, discharge summaries of admissions in July 2019 were exported from the EHR system.

A Venn diagram was used to visualize the overlap of identification by the two methods. In a qualitative analysis, cases with identification by ICD-10 codes but without identification in discharge summaries were examined. In addition, frequencies of each identification method were analysed using descriptive statistics for each hospital separately.

For further analyses, the identified hospitalisations were split into two groups:

- Group *ICD-10 code*: hospitalisations with an ICD-10 code for delirium (F05 or F10.4) recorded in the EHR system
- Group *Discharge summaries only*: hospitalisations without assigned ICD-10 code for delirium but with indication of delirium in discharge summaries

Descriptive statistics were used to describe the delirium patients of each group. Differences between the two groups were tested using a Mann-Whitney *U*-Test for ordinal variables and a Chi-squared test for nominal variables. A Bonferroni correction was used to correct for multiple testing resulting in an alpha-level of 0.008 for six variables. In order to determine differences in the discriminative performance, AUROC values and ROC plots with 95%-CIs were analysed for each group.

For the exploratory analysis of text patterns, a text mining method based on the *tm* package in R was used (Feinerer & Hornik, 2020). In a first step, discharge summaries of each hospitalisation were joined into a text vector. Patient names, numbers and special characters were removed from the texts. German stop words defined by the *tm* package and use case specific stop words, which were iteratively collected, were removed from the texts. The list of use case specific stop words is included in Table B.4 in the Appendix. Multiple occurrences of words within each text were removed, resulting in a unique vector of words for each hospitalisation.

Two exploratory analyses were performed on the text vectors of each group. First, the thirty most frequent words of both groups were compared using a bar chart with relative frequencies. Second, all texts were searched for words pertaining to one of the following categories: delirium, medication for delirium, confusion, disturbance, alcohol and dementia. Finally, relative frequencies of occurrences within each group were compared.

7.3 RESULTS

7.3.1 *Descriptive Statistics*

Over the year 2019, delirium risk was predicted for 19,050 admissions in five KAGes hospitals. After the exclusion of 1,087 admissions with predictions in July 2019, the cohort for analysis resulted in 17,963 admissions. Descriptive statistics for the cohort of each hospital are presented in Table 7.1.

The surgical department of Hospital 4 treated the youngest patients with a median age of 69 years. Most male patients were present in the cohort of Hospital 5, the cardiology department, and most female patients in Hospital 3. While the frequency of a past delirium recorded in the patient history was highest for Hospital 3 (2.7%), it was lowest for Hospital 5 (1.5%). Hospital 1 treated the patients with longest hospital stays (*median* = 6 days; Q1-Q3: 3.1-10.2) and with highest mortality (3.4%).

Hospital 2 includes two hospital locations in Styria, and one of them hosts a geriatric rehabilitation ward. At this ward, treated patients were at an older age (*median* = 78; Q1-Q3: 65-85) and had long hospital stays (*median* = 7 days; Q1-Q3: 4.0-14.1).

7.3.2 *Prospective Long-Term Performance*

Table 7.2 presents the results of the prospective performance for each hospital.

Table 7.1: Descriptive statistics of admissions from five KAGes hospitals with prospective risk predictions of delirium in 2019 (n = 17,963).

		Hospital 1		Hospital 2		Hospital 3		Hospital 4		Hospital 5	
N		8,727		3,988		3,624		1,017		607	
Age[a]	**Median**	72		72		73		69		71	
	Q1-Q3	58-81		57-82		57-81		55-79		58-79	
LOS[b]	**Median**	6.1		5.2		4.3		3.7		5.1	
	Q1-Q3	3.11-10.15		3.01-8.89		2.27-8.15		2.24-8.08		2.23-8.56	
		n	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
Sex	**male**	4,394	50.3	1,972	49.4	1,656	45.7	589	57.9	386	63.6
	female	4,333	49.7	2,016	50.6	1,968	54.3	428	42.1	221	36.4
Past Delirium		177	2.0	98	2.5	98	2.7	15	1.5	5	0.8
Mortality		298	3.4	113	2.8	57	1.6	9	0.9	6	1.0
Dep.[c]	**Surgical**	2,661	30.5	1,452	36.4	1,333	36.8	1,017	100.0	-	-
	Internal	6,066	69.5	2,536	63.6	2,291	63.2	-	-	607	100.0

[a] in years; [b] Length of stay, in days; [c] Department

Most predictions were evaluated for Hospital 1 (n = 8,727) with the longest period of deployment, and fewest predictions for Hospital 5 (n = 607) with a deployment period of three months only. The relative frequency of recorded delirium was highest for Hospital 5 (3.8%) and lowest for Hospital 4 (1.5%).

The distribution of patients stratified to each of the three risk groups varied slightly between the hospitals. Between 78.0% and 84.4% of admissions were stratified to the *low risk* group and between 8.0% and 12.8% to the *high risk* group. In Hospital 4, 13.2% of the admissions were stratified to the *very high risk* group, while in Hospital 5 only 5.4% were included in this group.

For Hospital 1, the distribution of the stratification changed after the update in July. From August until December, 2019, more patients were stratified to the *high risk* or *very high risk* group than from January until June.

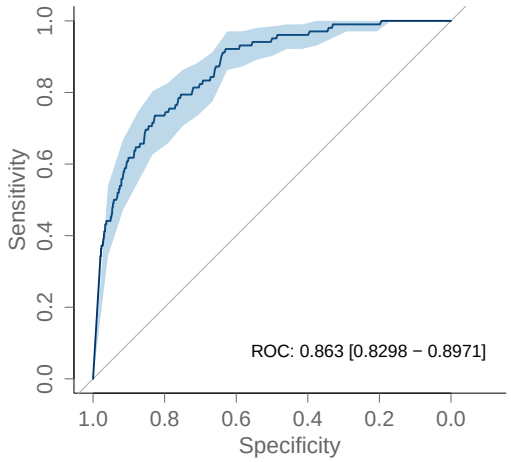
ROC plots in Fig. 7.1 complement the results of the discriminative performance demonstrated in Table 7.2. The highest discrimination for the last prediction was observed for Hospital 1 and the first version of the algorithm with an AUROC of 0.86 [0.830 - 0.897], followed by an AUROC of 0.83 [0.788 - 0.873] for Hospital 3 with the updated version of the algorithm. For Hospital 1, Hospital 2 and Hospital 3, AUROC values for the last prediction were above 0.80, while for Hospital 4 and Hospital 5 the algorithm achieved a lower performance with AUROC values above 0.76 only.

However, the ranges of the CIs limit the interpretation of the results for Hospital 4 and Hospital 5.

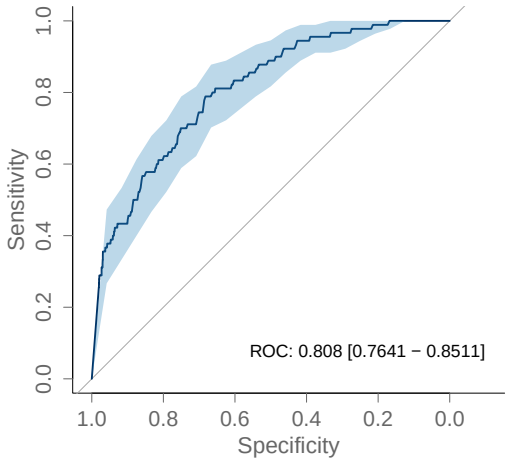
The discriminative performance increased from the first prediction to the last prediction for all hospitals except Hospital 4. However, CIs for AUROCs of the first and last prediction overlapped for all hospitals and thus the interpretation of these changes is limited.

For Hospital 1, the first version of the algorithm applied during the first half of the year 2019 achieved a better discriminative performance; the AUROC value decreased from 0.86 to 0.81 when using the updated version of the algorithm for the second half of the year.

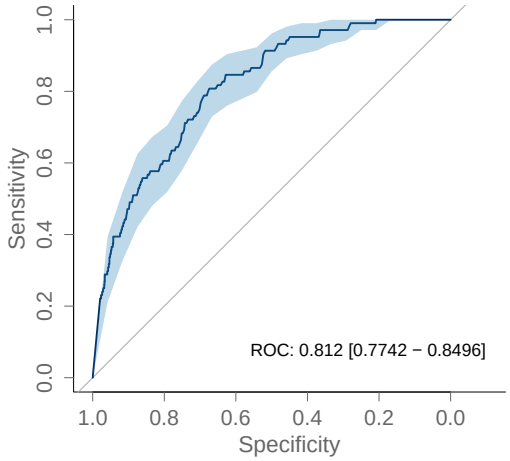
Calibration was poor for all hospitals, indicating an overestimation of the risk for all percentiles of predicted probabilities. Calibration plots for each hospital are included in Fig. A.7 in the Appendix. In concordance with the results of the discriminative performance, CIs for calibration curves were largest for Hospital 4 and 5.



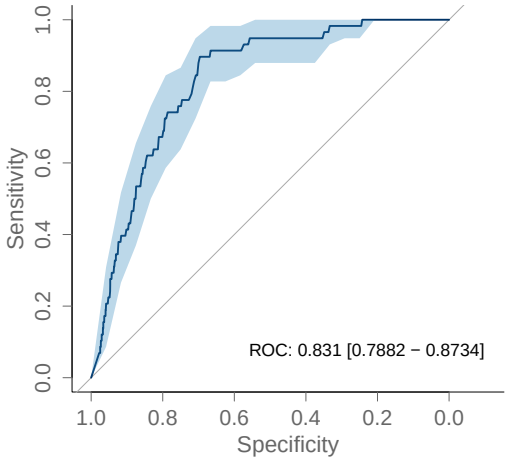
(a) Hospital 1 (Jan.-June)



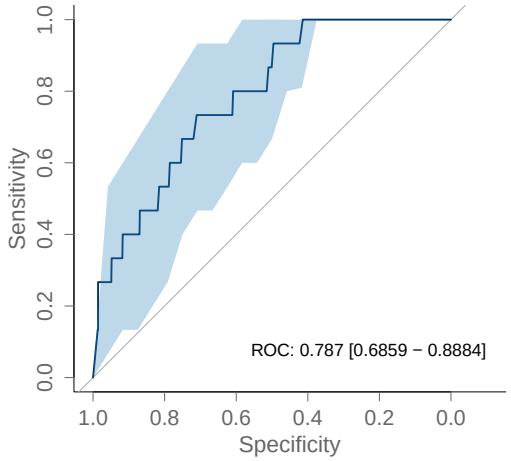
(b) Hospital 1 (Aug.-Dec.)



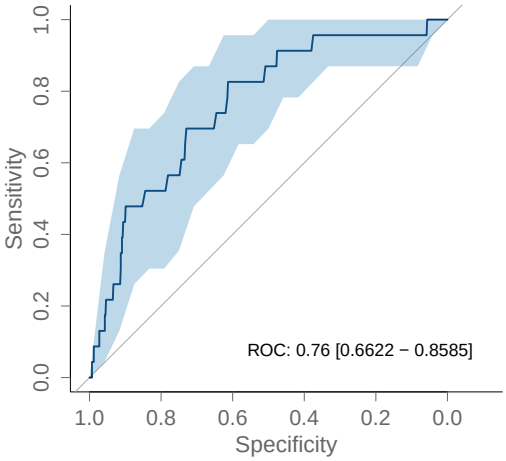
(c) Hospital 2 (Aug.-Dec.)



(d) Hospital 3 (Aug.-Dec.)



(e) Hospital 4 (Sep.-Dec.)



(f) Hospital 5 (Oct.-Dec.)

Figure 7.1: ROC plots of the last prediction within 48 hours of hospital stays in 2019. Hospital 1 used the first version (a) and the updated version of the algorithm (b).

Table 7.2: Prospective prediction results for five hospitals in 2019. Distributions for the three risk groups are presented using realtive frequencies. AUROC values including 95%-CIs are povided for the first prediction and the last risk prediction within 48 hours of hospitalisation.

Hospital	**Algorithm Version**	**Time period (2019)**		**N**	**Delirium**		**Risk group frequency (%)**			**AUROC [95%-CI]**			
		Start	End		n	%	low	high	very high	First prediction		Last prediction	
1	First	01.01.	30.06.	4,912	102	2.08	84.0	8.5	7.5	0.850	[0.8147 - 0.8861]	0.863	[0.8298 - 0.8971]
	Updated	01.08.	31.12.	3,815	90	2.36	78.0	12.8	9.2	0.798	[0.7552 - 0.8402]	0.808	[0.7641 - 0.8511]
2	Updated	01.08.	31.12.	3,988	104	2.61	81.8	10.1	8.2	0.785	[0.7446 - 0.8255]	0.812	[0.7742 - 0.8496]
3	Updated	01.08.	31.12.	3,624	58	1.60	80.0	11.1	8.9	0.788	[0.7357 - 0.8399]	0.831	[0.7882 - 0.8734]
4	Updated	15.09.	31.12.	1,017	15	1.47	78.9	8.0	13.2	0.832	[0.7434 - 0.9206]	0.787	[0.6859 - 0.8884]
5	Updated	03.10.	31.12.	607	23	3.79	84.4	10.2	5.4	0.702	[0.6010 - 0.8023]	0.760	[0.6622 - 0.8585]

7.3.3 *Exploratory Analysis of Feedback from Healthcare Professionals*

From August until December 2019, healthcare professionals provided feedback for 32 patients through the web application of the delirium risk stratification tool. The majority of the feedback regarded patients stratified to the *very high risk* group (n = 17). For five patients, healthcare professionals did not report a rating of the delirium risk, but provided a free-text comment.

Table 7.3 compares the risk ratings by healthcare professionals with the risk stratification by the algorithm. For 17 out of 32 patients (53.1%), healthcare professionals verified the risk stratification of the algorithm, indicated by bold numbers in the table.

For ten patients (31.2%), six with risk ratings and four with comments only, the delirium risk predicted by the algorithm was perceived as too low by healthcare professionals. Three patients were predicted a *high risk* by the algorithm, but healthcare professionals rated their risk as *very high*. For seven out of these ten patients, healthcare professionals commented the presence of delirium or confusion in the free-text field.

For four patients (12.5%), the delirium risk predicted by the algorithm was perceived as too high. Two of them were predicted a *very high risk* by the algorithm, and the other two a *high risk*. For one patient, the head of the cardiology department reported that the elevated liver enzymes were characteristic for the patient's infarct, and that the very high risk of delirium could not be confirmed from a clinical perspective. The remaining three patients had no risk factors and showed no signs of delirium during their hospital stay according to the healthcare professionals.

For the last patient with comment only, the healthcare professional reported that the *very high risk* prediction by the algorithm was not transparent enough. However, he or she provided no clinical risk rating nor further details.

Table 7.3: Feedback from healthcare professionals compared with the risk stratification of the algorithm. The feedback included a risk rating and/or comments.

		Risk stratification (Algorithm)			
		Low risk	High risk	Very high risk	Total
Risk rating	Low risk	**1**	2	2	5
	High risk	2	**2**	0	4
	Very high risk	1	3	**14**	18
Comments only		3	1	1	5
	Total	7	8	17	32

7.3.4 *Analysis of Discharge Summaries of Delirium Patients*

ICD-10 codes for delirium were recorded in the EHR system of the hospital network for 153 hospitalisations of the five hospital.

The text mining of discharge summaries identified 448 hospitalisations with indication of delirium. A manual check of these notes verified delirium corresponding to the ICD-10 code F05 for 388 hospitalisations (86.6%), and delirium corresponding to F10.4 for 21 hospitalisations (4.7%). Discharge summaries of 32 hospitalisations (7.1%) indicated a past delirium in the patient history; for the remaining 7 hospitalisations (1.6%), delirium could not be verified. This resulted in 409 hospitalisations with delirium identified by the text mining method.

The Venn diagram in Fig. 7.2 demonstrates the overlap of delirium identification using the two methods. Out of 153 hospitalisations with an ICD-10 code for delirium assigned, 138 (90%) were also identified by the text mining method. Frequencies of delirium identification for each hospital separately are included in Table B.3 in the Appendix.

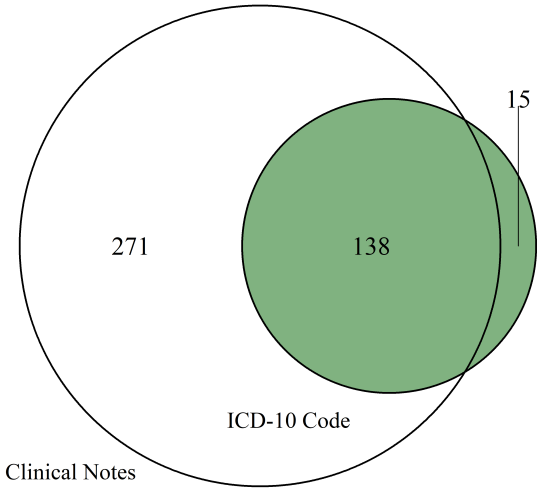


Figure 7.2: Venn diagram illustrating the overlap of ICD-10 code assigned delirium and indication of delirium in discharge summaries of the EHR system.

15 hospitalisations with an assigned ICD-10 code of delirium were not identified by the text mining method, and were further examined.

- Two patients were coded with F05 including the abbreviation *St.p.* in the diagnosis text. The abbreviation stands for *status post* and describes a patient's condition after the occurrence of a disease. This indicates that delirium did not occur in the current but in a previous hospital stay.

- One patient had a transfer between two KAGes hospitals during the hospital stay. Clinical texts before the transfer included indications of delirium, but discharge summaries described a delirium *status post* only. The patient was incorrectly flagged with no indication of current delirium during the manual check.
- Discharge summaries of seven patients indicated the occurrence of delirium by the following words: *disorientation*, *psychomotor disturbance*, *agitation*, *night-time confusion* and *recurring confusional state*. However, these words were not yet included in the list of search terms for text mining.
- For two patients, delirium was coded by anaesthesiologists at the ICU. The daily documentation of the ICU included records of agitation and delirium for one patient and records of sedation for the other patient. However, this document was not part of the exported clinical texts.
- One patient had no indications of delirium in the discharge summary, but the summary included the word *Temesta*, a drug used to treat agitation and disturbance.
- For one patient, F05 was coded together with the diagnosis text *Senile Dementia*. Besides the ICD-10 code, no record available in the EHR system indicated the occurrence of delirium, suggesting a coding error for this patient.
- For the last patient, discharge summaries were not exported from the EHR system because of technical issues.

For eight of the 15 admissions, indications of delirium were found in the daily nursing notes. These notes, just like the daily ICU documentation, were not yet included in the export of clinical texts.

For further analyses, hospitalisations were split into the two groups described in Section 7.2.5. The group *ICD-10 code* included 153 hospitalisations with delirium coded by F05 or F10.4, illustrated in green in Fig. 7.2. The group *Discharge summaries only* included 271 hospitalisations, illustrated in white.

The group *Discharge summaries only* had a significant lower median LOS with nine days than the group *ICD-10 code* with twelve days (see Table 7.4). There were no significant differences between the two groups in age, sex, history of delirium, mortality and specialization of the department. As shown in Fig. 7.3, the discriminative performance of the algorithm was better for the *ICD-10 code* group (AUROC = 0.868) than for the *Discharge summaries only* group (AUROC = 0.796).

The text mining method identified the 30 most frequent words from the discharge summaries of each group. The combination of these 30 words and their relative frequencies among each group are presented in Fig. 7.4. The word *delir*, German for delirium, occurred in 75.6% of the discharge summaries of the group *Discharge*

Table 7.4: Descriptive statistics for hospitalisations with ICD-10 coded delirium (n=153) or indication of delirium in discharge summaries only (n=271).

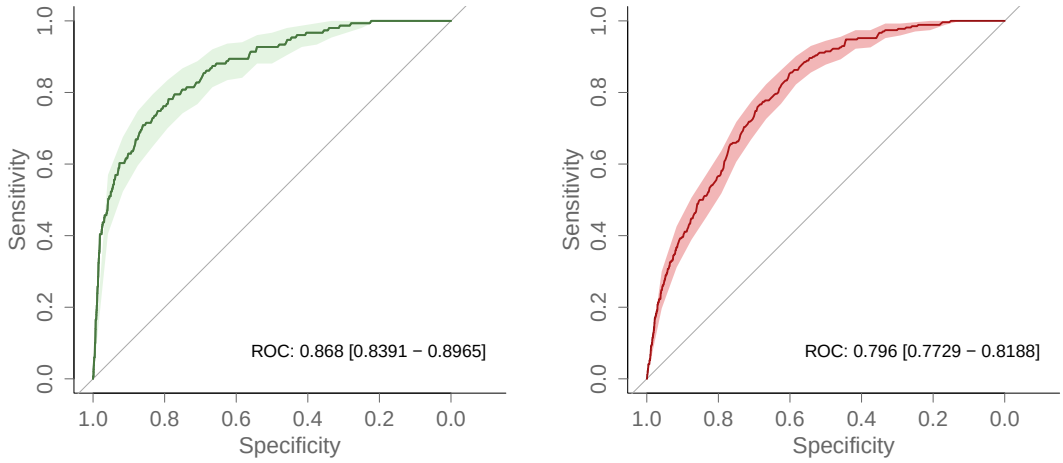
		ICD-10 code		Summaries only		
		Median	**Q1-Q3**	**Median**	**Q1-Q3**	***P* value (*U*)**
Age, years		82	77-89	81	75-88	0.0880
LOS, days		12	6.8-20.5	9	5.8-15.3	0.0062
		n	**%**	**n**	**%**	***P* value (χ^2)**
Sex	male	100	65.4	163	60.2	0.3382
	female	53	34.6	108	39.9	
Past Delirium		0	0.0	1	0.4	-
Mortality		31	20.3	36	13.3	0.0796
Department	Surgical	22	14.4	47	17.3	0.5111
	Internal	131	85.6	224	82.7	

summaries only, and in 86.9% of the group *ICD-10 code*. Besides delirium, the top five most frequent words included the German words for dementia, hypertension, chest X-ray and pleural effusion, followed by antibiosis, oriented, chronic, atrial fibrillation and disturbance.

Table 7.5 presents the search words related to six delirium relevant categories. Words related to all categories except the category *alcohol* were more frequent in discharge summaries of the *ICD-10 code* group.

Table 7.5: Relative frequency of delirium relevant words for two groups of delirium patients. Discharge summaries of both groups were screened for any search term of the category.

		Relative frequency	
Category	**Search terms (in German)**	**ICD-10 code**	**Summaries only**
Delirium	delir*, alkoholentzugsdelir*	0.876	0.793
Medication	seroquel, risperdal, risperidon, temesta, lorazepam, quetiapin, haldol	0.830	0.624
Confusion	verwirr*	0.438	0.362
Disturbance	unruhe, unruhig, unruhezust*, agitiert, agitation	0.549	0.406
Alcohol	alkohol*	0.288	0.299
Dementia	demenz, dement, alzheimer	0.628	0.443



(a) ICD-10 code (b) Discharge summaries only

Figure 7.3: ROC curves for the two groups of delirium patients stratified by the availability of ICD-10 codes in the EHR system.

7.4 LIMITATIONS

The aim of the third part of the thesis was to gain knowledge of the long-term performance of the delirium risk stratification tool when deployed across different hospitals within the KAGes hospital network. Several limitations need to be addressed regarding the study design, measures of performance and the data used for analysis.

7.4.1 *Limitations of the Study Design*

The evaluation periods varied between the hospitals. In Hospital 1, the tool had been used over the entire year 2019. Because the algorithm was updated in July 2019, the performance analysis had to be split. This resulted in an evaluation period of six months for the first version and five months for the updated version of the algorithm. The evaluation periods for Hospital 4 and Hospital 5 were shorter, resulting in wider CIs for performance measures.

The discriminative performance in Hospital 1 decreased from an AUROC of 0.86 to 0.81 when switching to the updated version (although with overlapping CIs). However, the study design did not allow for a proper investigation of this change. The incidence of delirium may underly seasonal variation (Balan et al., 2001), which limits the comparison of the results of the two versions.

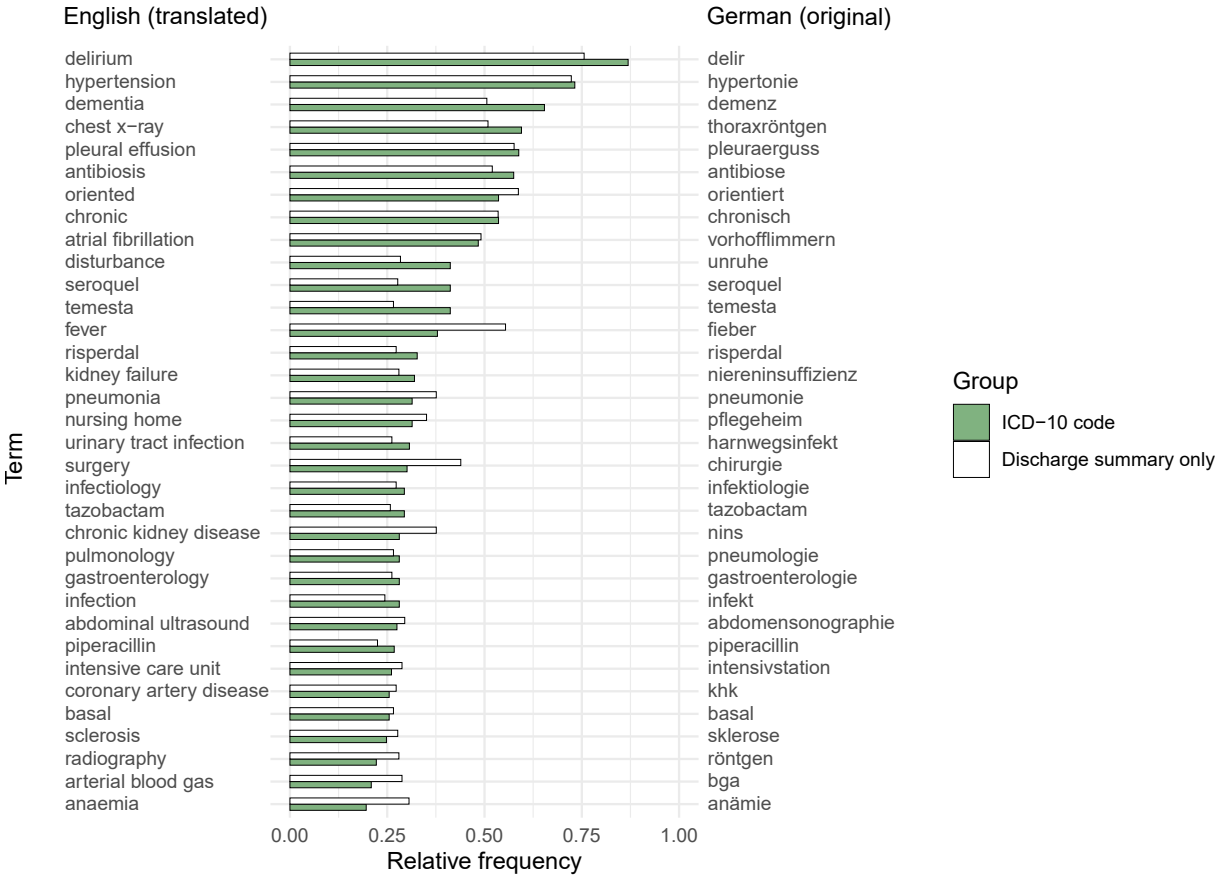


Figure 7.4: Bar chart illustrating the most frequent words in discharge summaries of delirium patients in English (translated) and German (original). Extracted words are illustrated in green for ICD-10 coded delirium patients, and in white for patients with delirium indicated in discharge summaries only.

7.4.2 *The Need for Alternative Performance Measures*

The CIs of the AUROC for Hospital 4 and Hospital 5 ranged from 0.69 to 0.89, and from 0.66 to 0.86, respectively. The interpretation of the discriminative performance is thus limited for both hospitals. However, during an assessment not included in this thesis, ward nurses from Hospital 4 reported a high accuracy of the algorithm. Although 21.2% of patients were stratified to the *high risk* or *very high risk* group, the ward nurses perceived the majority of them as correct. In contrast, senior physicians from Hospital 5 perceived the performance of the algorithm as rather poor. This highlights the need for alternative methods of evaluation rather than relying on AUROC and calibration plots only.

7.4.3 *Limitations of Clinical Data*

The final limitations regard the analysed data in this study. First, incidence of delirium was low, ranging from 1.5 to 3.8%. In the qualitative analysis, EHR data from patients with delirium-related ICD-10 diagnoses but without identification by text mining were examined. For some cases, indications of delirium were found in documents which were not yet included in the data used for text mining, e.g. ICU documentation or nursing notes. This implies that some delirium patients still remain unidentified.

Second, the analysis of discharge summaries of delirium patients may underlie a selection bias. Use case specific stop words for text cleansing and delirium relevant word categories were selected in an iterative process, which depended highly on the underlying data.

Third, the feedback received from healthcare professionals is limited to one aspect. The number of reports with an over- or underestimation of risk may be biased if users are more eager to report feedback for patients with incorrect prediction results. Nevertheless, the possibility to directly report feedback remains an important functionality of the tool.

8

DISCUSSION

In this thesis, a machine learning-based algorithm predicting delirium in hospitalised patients was evaluated during its use in clinical routine. The evaluation addressed three main research questions: (1) the performance of the algorithm in a clinical setting, (2) healthcare professionals' acceptance of the tool integrated into the HIS, and (3) the algorithm's long-term performance in a multicenter setting.

This chapter shortly presents the main results, discusses limitations of the evaluation, addresses advantages and disadvantages of machine learning-based prediction models, and provides an outlook for future research.

8.1 MAIN FINDINGS OF THIS THESIS

8.1.1 *Performance of the Delirium Risk Stratification Algorithm*

In the first part of this thesis, the performance of the delirium risk stratification algorithm was evaluated during a pilot study of seven months (Jauk et al., 2020). Delirium risk was predicted for 5,530 admissions within the first 24 hours of hospital stay using a random forest-based algorithm. In the prospective evaluation, the discriminative performance of the algorithm was excellent with an AUROC of 0.855 [0.8146-0.8956], a specificity of 81.8% and a sensitivity of 75.3%. In contrast, calibration of the algorithm was poor, which was most likely affected by a low number of patients with identified records of delirium (1.5%).

A re-training of the models and update of the algorithm in July 2019 addressed inaccurate predictions, which occurred for patients with systematically missing predictors within the first hours of the hospital stay. A simulation study on KAGes data demonstrated a better performance when adapting the models to the data available, e.g. for lab data and nursing data (Jauk et al., 2019). This corroborated the training of different models for three points of time of prediction accounting for the data available at each point of time.

However, results of the long-term evaluation in the third part of the thesis showed no improvement in AUROC for the updated version of the algorithm. Contrary, AUROC values were lower for the updated version than for the first version. This comparison

is limited by overlapping CIs and different months of evaluation, and further research is needed to explain the results and to verify this performance decrease.

During the long-term evaluation, the AUROC of the updated algorithm ranged from 0.76 to 0.83 across five hospitals. For the cardiology department in Hospital 5, the performance of the algorithm was not satisfying. With an AUROC of 0.76 discrimination was only acceptable, and internally collected feedback of healthcare professionals revealed a low perceived accuracy. Future studies are needed to evaluate the benefit of a prediction model trained exclusively on cardiac patients in order achieve a better performance.

Several delirium prediction models have been published, and some were included in the review in Section 2.5. The best performing model was a random forest model from Corradi et al. (2018), which achieved an AUROC of 0.86 on validation data. A GBM model developed by Wong et al. (2018) achieved an AUROC of 0.86 predicting delirium on validation data. At a specificity set at 90%, the sensitivity of the model was 59.7%. However, none of these models have been evaluated in clinical routine.

Compared to the reviewed delirium prediction models, the delirium risk stratification algorithm achieved a satisfying discriminative performance in the clinical setting. The random forest models included in the algorithm achieved AUROCs of 0.85 and 0.93 when evaluated on the test data, decreasing to values between 0.76 and 0.86 in prospective clinical settings.

8.1.2 *Validation by Clinical Experts*

In addition to the analysis of the prospective performance, a second method of evaluation was used in the first part of the thesis. Risk predictions of the algorithm were compared with the risk ratings of experienced ward nurses. The comparison was performed twice: in a blinded setting before the first implementation of the delirium risk stratification tool, and in a non-blinded setting with the tool implemented.

During two weeks of the first comparison, the prediction was not yet visible in the HIS, but the algorithm was already running in the background. This comparison provided a first quality assessment for the implementation in clinical routine before making the tool available for healthcare professionals.

In both comparisons, the ratings of the ward nurses and the prediction of the algorithm were in agreement for the majority of the cases. The qualitative analysis revealed not only a known weakness of the algorithm, e.g. inaccurate risk when little data was available, but also its strengths. The application provides a good support in delirium management if healthcare professionals are uncertain or if patients are unable

to communicate. For instance, patients under sedation or with language barriers might benefit from the interaction of healthcare professionals with the delirium risk stratification tool.

Similarly, Brennan et al. (2019) compared physician's risk assessment of postoperative complications with a machine learning-based algorithm. The algorithm outperformed physicians' assessments in five out of six complications. After their interaction with the algorithm, physicians' assessments improved for two use cases. In this thesis, an evaluation of changes in the decision-making of healthcare professionals was not possible because of the low incidence of delirium in both comparisons.

Over the entire deployment process, feedback had been collected from physicians and nurses involved in an expert group (described in Section 3.2). Especially the discussion with experts about the distribution of risk groups was highly relevant for the clinical success of the tool. Based on the feedback from the experts, the distribution has been constantly evaluated and adapted for the departments since the first deployment in 2018.

However, it remains a challenge to set optimal thresholds in order to separate the risk groups. For some patients, risk probabilities are right at the boundary to the other risk group. In the web application, a bar chart facilitates the identification of patients close to another risk group. This supports the adaptation of delirium management for patients at the very lower or very upper limit of a risk group.

After the update of the algorithm in July 2019, the validation by clinical experts was further enhanced by adding a feedback button to the web application. During the long-term evaluation in the third part of the thesis, feedback had been collected for all participating hospitals through this button.

One feedback highlighted a problem caused by the random forest model predicting alcohol withdrawal delirium (F10.4). For patients with alcohol abuse, laboratory values indicating liver problems like alanine aminotransferase, aspartate aminotransferase or gamma-glutamyltransferase are typically out of norm. However, also for patients with pancreatitis, viral hepatitis or cardiovascular disease such values are often out of norm, which resulted in a stratification to the very high risk group for several patients. The F10.4 model might presume alcohol abuse for the patient – a risk factor of delirium – and misleadingly predicts a higher risk of delirium.

Overall, feedback by healthcare professionals is essential to reveal potential weaknesses of tools in clinical routine. It helps to better understand how healthcare professionals perceive machine learning-based tools and which problems may arise during their use in clinical routine.

Based on the feedback, patients with risk predictions contradictory to the ratings of healthcare professionals can be better presented in future model training. This might further improve the performance of the delirium risk stratification algorithm in clinical settings.

8.1.3 *Technology Acceptance of the Delirium Risk Stratification Tool*

The second part of the thesis addressed the technology acceptance of the delirium risk stratification tool from a user-centric perspective (Jauk et al., 2021). The well-established TAM (Davis, 1989) and parts of its extension TAM2 (Venkatesh & Davis, 2000) were used to frame the evaluation process, assessing *perceived usefulness*, *perceived ease of use*, *output quality* and *actual system use*.

Quantitative and qualitative methods were used in a convergent parallel study design. Besides conducting an anonymous questionnaire assessment involving 10 physicians and 37 nurses, clinicians from the expert group provided regular feedback, which was analysed qualitatively. The expert group supported the improvement of visualization and algorithm functionality over the entire deployment process to a great extent.

After seven months of deployment in clinical routine, the majority of users reported that the tool was useful for the prevention of delirium and its early detection. According to the experts, it offered a great support and helped to reduce resources needed for delirium screening. Healthcare professionals reported that the tool supported delirium management especially for patients with delirium diagnosed in the past, who are automatically stratified to the high risk group.

Users and clinical experts appreciated the visualization of the risk stratification results in the HIS and the detailed summary of the risk prediction in a web application. The automatic and fast prediction without the need of manual data entry presented a great value to them.

However, not all users were able to integrate the tool in their daily clinical practice and not all of them felt sufficiently prepared to use it. This is in line with the results on actual system use, which was reported to be low. Only 28% of the users reported that they had been using the tool regularly, and the expert group concluded that more promotion and more training sessions were needed in future.

Research on implemented machine learning models is rare, and few studies have evaluated user acceptance and technology uptake in such contexts. A study by Brennan et al. (2019) assessed the user acceptance of a machine learning application for preoperative risk assessment. Their study sample was small and homogeneous,

including ten physicians only, and feedback was rather negative. Only half of the physicians reported that they would use the application and that it helped them in decision-making.

Also in this evaluation study, only ten physicians reported feedback for the tool. However, in contrast to the study by Brennan et al. (2019), nurses were also included in the quantitative assessment, because the assessment of delirium risk is highly relevant for them. This raised the number of participants to 47. However, a comparison of technology acceptance between nurses and physicians was not possible due to the group imbalances.

Ginestra et al. (2019) evaluated the perception of a machine learning algorithm predicting severe sepsis and septic shock. The evaluation of the algorithm, which achieved a sensitivity of 26% and a specificity of 98% in a validation cohort, included a large, heterogeneous sample of different user groups. The perception of nurses and physicians differed substantially in the evaluation. Physicians evaluated the tool rather poorly due to missing transparency and late alerts. Only a fifth of the physicians found the alerts helpful, in contrast to almost half of the nurses. Furthermore, nurses reported an improvement of care due to the alerts more often than physicians.

These results indicate that the benefit of risk prediction may differ between groups of healthcare professionals. Future evaluation of the delirium risk stratification tool should therefore strive for results divided by user groups when evaluating the technology acceptance.

8.2 RECORDS OF DELIRIUM IN EHR SYSTEMS

8.2.1 *Under-Diagnosing of Delirium*

A major obstacle for the development of delirium prediction models is the diagnosis of delirium itself. There are no laboratory or imaging tests to show the presence of delirium in patients. Different severities and different subtypes of delirium lead to a broad range of signs and symptoms, which further complicates the diagnosis. Even experienced clinicians may have difficulties to distinguish delirium from dementia in clinical settings (Fong et al., 2015).

Difficulties in diagnosing delirium lead to many undiagnosed cases, especially for patients with a history of dementia (Lange et al., 2019). This underdiagnosis is one reason why delirium in under-recorded in EHR systems. Especially hypoactive delirium is prone to systematic under-diagnosing, which may limit the development of prediction models. If there are less records for hypoactive subtypes available in

the EHR system, models may learn clinically incorrect patterns and will misclassify hypoactive delirium patients in the future.

The incidence of delirium observed in the studies of this thesis indicates that delirium is also under-recorded in the EHR data of KAGes. The highest incidence was assessed at the cardiology department of the university hospital with 3.8% delirium patients during three months. For data of the pilot study, the incidence had been only 1.5% over seven months.

Several factors might have led to this low number of delirium patients identified in the clinical setting:

- A high risk of delirium may be correctly predicted by the algorithm and the occurrence successfully prevented due to clinical interventions
- Mild forms of delirium may be documented less often when compared to severe forms
- Hypoactive delirium may be overseen and not documented
- Physicians not specialized in the fields of neurology or psychiatry may avoid coding of delirium because of unclear diagnostic criteria

The low observed incidence did not only influence the evaluation of the tool, it also raises the question how to better identify delirium in EHR data.

8.2.2 *Effects of a Low Incidence for the Predicted Outcome*

The observed low number of delirium patients limits the interpretation of the results of this thesis to a certain extent. Measures of discrimination derived from the confusion matrix, such as sensitivity or positive predictive value, are strongly influenced by under-recording. Not identified delirium patients are incorrectly included in the non-delirium group for analyses.

Calibration plots of the test data showed a good calibration for the random forest models, but plots of the prospective prediction illustrated an overestimation of delirium risk for almost all percentiles and cohorts. When assuming a high rate of under-recording, the results of the calibration plots are not informative.

Even though the observed low incidence had a major impact on the evaluation, its impact on model development is minor (Jauk, 2021). For training the random forest models, patients with corresponding ICD-10 codes (either F05 with subcategories or F10.4) were labelled as delirium. Internal analyses showed that coding errors occurred for very few of these cases, and it can be assumed that the cohort of delirium patients was accurate for training.

In contrast, under-recording results in delirium patients incorrectly labelled as controls. In order to identify such false negative cases, discharge summaries were screened and manually checked for mentions of delirium-related language expressions. Positive cases were excluded from the control group, which resulted in controls without ICD-10 codes nor mentions of delirium in discharge summaries and led to a quite accurate group of controls.

A study by Ryan et al. (2013) identified severity of inattention and being under medical care as predictors of a complete delirium documentation in case notes. The exploratory analysis in Section 7.3.4 showed no differences in delirium coding between surgical and medical departments, but delirium patients with ICD-10 coded delirium had significantly longer hospitals stays than delirium patients with mention of delirium in discharge summaries only. Furthermore, medication typical for delirium was mentioned more frequently in discharge summaries of ICD-10 coded patients. Longer hospitals stays and mentions of certain drugs could indicate higher severity, which suggests that delirium is coded more often in more severe delirium cases.

The delirium risk stratification algorithm achieved a higher discriminative performance when including patients with ICD-10 coded delirium only. One explanation could be that the algorithm predicts severe cases of delirium with higher accuracy than milder forms of the syndrome. Further research is needed in order to examine the hypotheses raised and to reveal more details of delirium coding patterns in the EHR system under scrutiny.

8.2.3 *Digital Phenotyping of Delirium Patients*

In response to the results published by Jauk et al. (2020), Rousseau & Tierney (2021) raised concerns about the identification of delirium patients, which depended mainly on ICD-10 coded diagnoses. The authors provided examples of additional data sources to improve the definition of clinical conditions, or so-called digital phenotyping, of delirium patients. According to them, lab values, vital signs or symptoms referenced in clinical texts may help to improve the positive predictive value.

Up to now, simple string matching methods have been applied to identify mentions of delirium in discharge summaries. Future work should investigate a potential benefit of using more sophisticated natural language processing (NLP) technology on these texts. This could improve the accuracy of the digital phenotype of delirium, and add new features to the prediction models. Structured information extracts from clinical texts can help to increase the performance, particularly for patients with little structured information available.

The exploratory analysis in the third part of the thesis demonstrated that sources apart from ICD-10 codes and discharge summaries could help to identify delirium patients. For the EHR system of KAGes, the following additional sources could refine the digital phenotype of delirium for future research:

- Psychiatric notes
- Drug dispensation
- Daily nursing notes
- Daily ICU notes

Although a broader definition of delirium can be useful for the evaluation, it can also be hazardous when prediction models are trained. Labelling clinical cases as delirium patients, with different subtypes, severity level or unspecific signs and symptoms, can result in inaccurate models. The use of delirium-relevant drug dispensation, for instance, can be too unspecific, as some of these drugs are also used to treat anxiety or sleeping disorders. Furthermore, electronic drug dispensation is not available across all KAGes hospitals, which introduces a bias in model training. Overall, a comprehensive evaluation and full understanding of the records used for digital phenotyping is crucial.

8.3 THE DEPLOYMENT OF MACHINE LEARNING-BASED ALGORITHMS

8.3.1 *Overcoming Barriers of Deployment*

Various barriers limit the deployment of new decision support systems using machine learning and EHR data, and several examples of these barriers are discussed in Section 2.2.1. Three unique features of the setting in KAGes helped to overcome these barriers and led to a successful deployment of the tool in various hospitals.

First, for all hospitals that belong to the KAGes network in Styria the same, shared EHR system has been in place for over 17 years. This facilitates the access to the amounts of data needed for training machine learning models as well as the deployment of the tool in various hospitals of the network.

Second, the tool has been internally developed by KAGes, which allows direct access and transfer of data needed for modelling, and facilitates continuous monitoring of the tool during deployment. The setting might furthermore increase the trust and acceptance of healthcare professionals, because their colleagues are involved in the development.

Third, due to this development from inside the organisation, close communication between healthcare professionals, EHR technicians and data scientists was maintained

during the whole development and deployment process, which enabled a successful technical integration of the tool. The close contact with nurses and physicians facilitated all steps of deployment and evaluation in the clinical setting, and initiated further use cases for prediction models.

8.3.2 *Evoked Changes in Healthcare*

The deployment of machine learning-based prediction models can induce many changes in healthcare systems. Wyatt & Spiegelhalter (1990) summarized several side effects that can influence decision support when deploying new systems.

One of these side effects is the Hawthorne effect, which describes people modifying their behaviour because of being the topic of an observation (Adair, 1984). Although the Hawthorne effect has been questioned and stimulated controversy (Franke & Kaul, 1978), Wyatt & Spiegelhalter (1990) assume that the performance of decision-makers improves just because their decisions are being studied. According to this assumption it is possible that healthcare professionals were more attentive for delirium during the study period.

This awareness might have further increased among the staff in the participating departments due to training sessions and the encouragement by heads of the department heads. This could have sharpened the awareness even across physicians and nurses from departments with lower incidence of delirium, and might have further increased the detection of delirium patients.

Another positive side effect of the deployment could be provoked by the extended use of EHR data for clinical purposes. Healthcare professionals invest a substantial amount of their time in EHR documentation. Demonstrating them a tool that reuses this information for clinical decision-making might raise the importance of documentation. For the delirium risk stratification tool, an already assigned ICD-10 code F05 in the EHR system had an immediate impact, stratifying these patients to the very high risk group in following predictions.

The deployment of a risk prediction model comes along with the need for preventive interventions. Potential extra workload, e.g. due to an extensive amount of alerts, has been identified as a barrier of implementation for decision support systems (Varonen et al., 2008). Although patients are not negatively affected by any of the non-pharmacological interventions, the clinical workload could rise unnecessarily with a too high number of false positive cases. Healthcare professionals reported that the use of the delirium risk stratification tool did not increase their workload, but future

evaluation should assess whether more clinical resources are needed when using the tool.

One effect of successful delirium prevention is discussed in Section 5.4.1: A self-destroying prophecy is caused for every patient with a predicted high delirium risk in which delirium is successfully prevented. Accordingly, Lenert, Matheny, & Walsh (2019) describe prediction models as 'victims of their own success'. As the purpose of prediction models is to improve clinical outcomes, successful models will decrease the distribution and variance of the predicted outcome. Systematic changes in the outcome can further lead to a decline in performance.

According to the authors, a simple re-training of the models without accounting for the changes in the clinical process behind will not prevent the misclassification of patients (Lenert, Matheny, & Walsh, 2019). The proposed way to deal with the problem is to model the intervention space and include the interventions into the model itself.

For the delirium case, non-pharmacological and multicomponent interventions are often not documented in detail. Although the modelling of such interventions presents various challenges, it is a promising method to enable the stability of machine learning-based prediction models over time.

8.4 STRENGTHS AND WEAKNESSES OF MACHINE LEARNING-BASED ALGORITHMS

Chapter 2 addresses several strengths and weaknesses of machine learning-based algorithms, and some of them need to be discussed regarding the results of this thesis.

8.4.1 *Weaknesses of the Delirium Risk Stratification Algorithm*

An often discussed weakness of complex machine learning models is their lack of interpretability. In order to avoid a black-box scenario, the delirium risk stratification tool includes a web application presenting a patient's EHR data used for prediction. In this thesis, users reported that the presented information was understandable and the application provided them with useful additional information.

As an attempt to explain the prediction, EHR data are ranked based on evidence-based risk factors from the literature and variable importance methods. However, for some cases it still remains a challenge to identify individual features which had a major impact on the results.

One limitation are the variable importance methods for random forest models, which are prone to be biased when variables vary in scale types and categories (Strobl et al., 2007). Results are misleading if suboptimal predictors are preferred artificially. This

may be induced by a biased feature selection in individual classification trees and by bootstrap sampling with replacement.

Furthermore, global variable importance is not able to provide the direction of correlation nor to explain individual prediction results. The use of local variable importance methods such as LIME (local interpretable model-agnostic explanations) instead of global variable importance can support the explanation of personalised risk prediction.

Enabling interpretability and transparency of machine learning models facilitates not only clinical decision-making and the appraisal of risk predictions, it can also reduce biases. The overreliance of users on clinical decision support systems, called automation bias, can lead to decision errors. A meta-analysis found that the errors increase up to 26% when automated systems give incorrect advice to users compared to not using decision support at all (Goddard, Roudsari, & Wyatt, 2012, cited by Lyell & Coiera, 2017). In order to identify errors of the system, verification of the results needs to be facilitated.

Besides automation bias, machine learning models are prone to be biased for other characteristics such as gender, ethnicity or demographics. One potential bias could be evoked by the prediction of alcohol withdrawal delirium for patients with elevated liver enzymes due to pancreatitis, hepatitis, or cardiovascular disease (discussed in Section 8.1.2). Future studies are needed to determine potential biases of the random forest models implemented in the delirium risk stratification algorithm.

Finally, the precise definition of the point of time of risk prediction needs to be discussed as a weakness of machine learning-based prediction. Contrary to established risk scores, for machine learning-based models it is crucial to know which predictors are available at the exact point of time in order to adapt the training data. Systematically missing values or an overlap of the point of time of prediction and predictors' availability can reduce the performance of prediction models when implemented in clinical settings.

Corradi et al. (2018) noted that not all factors that would be predictive for delirium were universally available for their delirium model. In addition, the authors used the information from the entire hospital stay for non-delirium patients in the training data set. This makes a risk prediction at admission infeasible in a clinical setting, because more information was available for model training than at the point of time of prediction during deployment. In the study by Kim et al. (2016), some predictors were collected after the possible onset of delirium, which also limits clinical deployment.

8.4.2 *Strengths of the Delirium Risk Stratification Algorithm*

Although several challenges need to be overcome when using machine learning models for risk prediction in healthcare, their advantages will predominate for various use cases in the future.

First, automated risk prediction based on EHR data can substantially reduce hospital resources. While most established risk scores require additional assessment and scoring, machine learning models are able to use available information for the risk prediction solely. No manual data entry or clinical assessment is needed to determine a patient's risk when using such tools.

Second, the direct integration of tools in the HIS enables a risk stratification within seconds after a patient is admitted. This immediate prediction is essential for a fast impression of a recently admitted patient in order to stage preventive actions as early as possible.

Third, the use of machine learning can support the identification of an actual risk of delirium rather than just the detection of the syndrome. As discussed in Section 2.4.3, screening instruments like DOS or CAM assess already existing signs of delirium. Although clinical protocols for delirium include predisposing or precipitating risk factors, machine learning models can account for many more parameters. For a syndrome or disease with a multifactorial onset like delirium, this is highly relevant.

Finally, risk prediction using machine learning can facilitate knowledge transfer. In the evaluation by Brennan et al. (2019), the risk assessment by physicians improved after their interaction with a machine learning algorithm. Even though senior healthcare professionals will mostly outperform machine learning algorithms within their field, inexperienced staff can learn from tools visualizing patients' risks and patterns of risk factors. More research is needed to determine a potential learning effect introduced by the delirium risk stratification tool.

8.5 THE CLINICAL BENEFIT OF DELIRIUM PREDICTION

The prediction of delirium presents an ideal use case for clinical decision support. First, it is important to identify patients at high risk because delirium is associated with higher mortality and morbidity of the patients. Second, its occurrence can be prevented by non-pharmacological interventions, which do not lead to any harm for patients. This facilitates the deployment of the tool, because even false positive cases are not harmfully affected by any preventive actions.

In hospital routine, the identification of high-risk patients presents a great challenge for healthcare professionals. Due to the multifactorial aetiology of delirium, individual risk factors are difficult to determine. In contrast to nursing homes, which treat patients in the long-term, hospital staff often lacks detailed information about patients and their individual risk factors.

The results of this thesis indicate that the delirium risk stratification tool is a reliable and effective support for the delirium management in hospitals. There are two more major advantages of the algorithmic approach compared to well-established screening methods. First, no additional information is needed to determine the delirium risk, because the algorithm uses already documented information only. Second, the algorithmic prediction is faster than standardised screening methods and does not depend on any clinical resources. Within the first minutes of a hospital stay, the tool automatically stratifies patients according to their delirium risk.

Nevertheless, the prediction results visualized in the HIS also raise questions how to proceed with patients at risk. There is still a lack of professional guidelines for a systematic management of delirium (Young et al., 2008). Delirium prevention needs to be person-centred and at high-quality, and this personalised prevention may be supported by machine learning methods in future.

8.6 OUTLOOK FOR FUTURE RESEARCH

Although this thesis covers important aspects of the evaluation of machine learning-based algorithms in clinical settings, several research questions remain open and should be addressed in future studies.

First, external validation of the delirium risk stratification tool in other hospital networks is needed. Although the algorithm achieved a satisfying performance when deployed in KAGes hospitals, its performance is likely to differ for hospitals with different EHR documentation and healthcare systems. The generalisability of machine learning models should be topic of further studies, including an evaluation of how much the models need to be adapted in order to achieve an acceptable performance. The promising technology federated learning helps to train models across various institutions without exchanging any data, which could lead to more generalisable models (Rieke et al., 2020).

Apart from adapting the models to new clinical settings, iterative model improvements will be necessary. On the one hand, stability of the performance over time should be obtained by re-training the models when needed. On the other hand, data used for modelling should be optimized. Further predictors can be created out of

information extracted from EHR narratives using NLP; digital phenotyping of delirium patients can be extended by additional methods of identification; and methods of feature selection and model training can be improved. For patients with few data available at a hospital, risk prediction could be improved by using additional data, e.g. by a nationwide EHR system such as the Austrian personalised health record system ELGA[1].

Future model improvements should also investigate the application of new learning and modelling methods. Further machine learning methods should be explored, especially when combining different sources of EHR data such as structured data, clinical texts or images.

Neural nets with a single hidden layer have been commonly used for clinical prediction models in healthcare, but the benefit of deeper neural networks with more than one hidden layer is still under investigation (Xiao, Choi, & Sun, 2018); long short-term memory (LSTM) models could increase the performance when modelling longitudinal patient histories (Lipton et al., 2016); transfer learning can help to overcome data scarcity (Desautels et al., 2017); generative adversarial networks (GAN) can provide plausible labelled EHR data and boost the performance when data are limited (Che et al., 2017); and automated machine learning (AutoML) simplifies the data preprocessing and training process, which allows for frequent re-training (Waring, Lindvall, & Umeton, 2020).

Even though new approaches in the field of artificial intelligence and machine learning are promising, methods should be chosen wisely, keeping in mind their interpretability, explainability and effort of deployment.

Regarding explainability, further improvements can be achieved for the delirium risk stratification tool in the future. Additional research on user interface and user experience can increase the benefit of the tool for healthcare professionals. Changes in usability and transparency need to be evaluated when applying methods such as LIME for an individual ranking of relevant predictors.

Finally, as stated several times, an on-going evaluation and monitoring of the delirium risk stratification tool will be necessary over the entire deployment period. This requires that data of the prospective predictions are constantly being collected and, whenever possible, feedback of users and clinical experts are gathered. Further evaluation studies using different designs and settings are needed to provide more information on the clinical impact of the tool. This includes an assessment of point prevalence of delirium, an evaluation of preventive actions or the use of randomized

1 https://www.elga.gv.at/en/about-elga/

study designs. In the coming years, different sources of information should guide decisions on how to proceed with the delirium risk stratification tool in clinical routine.

9 CONCLUSION

This thesis is among the first ones to evaluate the performance and acceptance of a machine learning-based risk stratification tool in clinical routine. Numerous machine learning models have been developed over the past years addressing highly relevant use cases in healthcare, but their performance in clinical routine has largely remained unknown.

The results of this work highlight the importance for evaluation of machine learning models in prospective clinical studies instead of reporting performance based on test data only. After an implementation in the HIS, the delirium risk stratification algorithm achieved a stable performance. However, both discrimination and calibration decreased slightly during prospective prediction when compared to the performance on the test data.

The results demonstrate two scenarios for which an on-going evaluation of machine learning models is indispensable. First, even though the continuous update and re-training of machine learning models is recommended in order to achieve stable performances, the results of this thesis illustrate that improvements in the test settings do not necessarily lead to an improvement in clinical settings. The updated version of the algorithm achieved a higher performance on the test data than the first version, but this performance gain could not be shown for the prospective prediction in clinical routine. As a consequence, prospective prediction results should be analysed with reasonable care after updating already implemented risk stratification tools with re-trained models.

Second, a decrease in performance was observed when using the delirium risk stratification algorithm on patients who were under-represented in the training population. Thus, in-depth monitoring of the clinical performance is essential, when applying machine learning models to patient cohorts that exhibit very distinct profiles compared to the respective training cohort.

Even though machine learning models achieve promising results on test data, their application in clinical routine might be useless. Hence, before starting the development of prediction models, the actual usefulness and potential clinical benefit should be verified by clinicians. First, the predicted outcome needs to be preventable and healthcare professionals have to be aware of the most effective preventive actions. The prediction of delirium for example does not lead to better care as long as there

is no proper awareness for its management in the respective wards. Moreover, a point of time of prediction needs to be set early enough in order to enable healthcare professionals to control the predicted outcome. For instance, predicting delirium in a late stage of a hospital stay might not allow for preventive actions any more.

The overall goal of machine learning-based risk stratification is to support healthcare professionals in their decision-making. In most studies, machine learning models are evaluated using measures for discrimination and calibration only. However, high discriminative performance and good calibration present only the basis for successful decision support. The most powerful way to determine whether a tool effectively supports decision-making in healthcare is to give voice to the actual users. Especially when it comes to explainability and interpretability of prediction results, opinions of clinical experts should be considered for the development of machine learning-based tools.

For a successful implementation in clinical routine, the acceptance of machine learning-based tools by healthcare professionals is an essential component. Without their belief in the usefulness, their support during the entire implementation process and regular feedback, the tool is doomed to failure. Only those machine learning tools that achieve high accuracy, predict actionable events and are highly accepted by healthcare professionals will be able to improve healthcare quality and hence patient safety in future.

BIBLIOGRAPHY

Adair, J. G. (1984). The Hawthorne Effect: A Reconsideration of the Methodological Artifact. *Journal of Applied Psychology* 69 (2), pp. 334–345.

Amarasingham, R., Patzer, R. E., Huesch, M., Nguyen, N. Q., & Xie, B. (2014). Implementing Electronic Health Care Predictive Analytics: Considerations And Challenges. *Health Affairs* 33 (7), pp. 1148–1154. ISSN: 0278-2715, 1544-5208. DOI: 10.1377/hlthaff.2014.0352.

American Psychiatric Association (1980). *Diagnostic and Statistical Manual of Mental Disorders*. 3rd Edition. Washington, D.C: American Psychiatric Association.

Ammenwerth, E. (2019). Technology Acceptance Models in Health Informatics: TAM and UTAUT. *Studies in Health Technology and Informatics* 263, pp. 64–71. DOI: 10.3233/SHTI190111.

Ancker, J. S., Senathirajah, Y., Kukafka, R., & Starren, J. B. (2006). Design Features of Graphs in Health Risk Communication: A Systematic Review. *Journal of the American Medical Informatics Association* 13 (6), pp. 608–618. ISSN: 1067-5027, 1527-974X. DOI: 10.1197/jamia.M2115.

Balan, S., Leibovitz, A., Freedman, L., Blagman, B., Ruth, M., Ady, S., & Habot, B. (2001). Seasonal Variation in the Incidence of Delirium among the Patients of a Geriatric Hospital. *Archives of Gerontology and Geriatrics* 33 (3), pp. 287–293. ISSN: 01674943. DOI: 10.1016/S0167-4943(01)00192-3.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients. *Health Affairs* 33 (7), pp. 1123–1131. ISSN: 0278-2715, 1544-5208. DOI: 10.1377/hlthaff.2014.0041.

Beam, A. L. & Kohane, I. S. (2018). Big Data and Machine Learning in Health Care. *JAMA* 319 (13), p. 1317. ISSN: 0098-7484. DOI: 10.1001/jama.2017.18391.

Benjamens, S., Dhunnoo, P., & Meskó, B. (2020). The State of Artificial Intelligence-Based FDA-Approved Medical Devices and Algorithms: An Online Database. *npj Digital Medicine* 3 (1), p. 118. ISSN: 2398-6352. DOI: 10.1038/s41746-020-00324-0.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., & Eckersley, P. (2020). Explainable Machine Learning in Deployment. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Barcelona Spain: ACM, pp. 648–657. ISBN: 978-1-4503-6936-7. DOI: 10.1145/3351095.3375624.

Bian, J., Buchan, I., Guo, Y., & Prosperi, M. (2019). Statistical Thinking, Machine Learning. *Journal of Clinical Epidemiology* 116, pp. 136–137. ISSN: 08954356. DOI: 10.1016/j.jclinepi.2019.08.003.

Bickel, H., Gradinger, R., Kochs, E., & Förstl, H. (2008). High Risk of Cognitive and Functional Decline after Postoperative Delirium. *Dementia and Geriatric Cognitive Disorders* 26 (1), pp. 26–31. ISSN: 1421-9824, 1420-8008. DOI: 10.1159/000140804.

Bihorac, A., Ozrazgat-Baslanti, T., Ebadi, A., Motaei, A., Madkour, M., Pardalos, P. M., Lipori, G., Hogan, W. R., Efron, P. A., Moore, F., Moldawer, L. L., Wang, D. Z., Hobson, C. E., Rashidi, P., Li, X., & Momcilovic, P. (2019). MySurgeryRisk: Development and Validation of a Machine-Learning Risk Algorithm for Major Complications and Death After Surgery. *Annals of Surgery* 269 (4), pp. 652–662. ISSN: 1528-1140. DOI: 10.1097/SLA.0000000000002706.

Breiman, L., Friedman, J., Stone, C., & Olshen, R. (1984). *Classification and Regression Trees*. Taylor & Francis. ISBN: 978-0-412-04841-8.

Breiman, L. (2001a). Random Forests. *Machine Learning* 45 (1), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324.

Breiman, L. (2001b). Statistical Modeling: The Two Cultures. *Statistical Science* 16, pp. 199–215.

Brennan, M., Puri, S., Ozrazgat-Baslanti, T., Feng, Z., Ruppert, M., Hashemighouchani, H., Momcilovic, P., Li, X., Wang, D. Z., & Bihorac, A. (2019). Comparing Clinical Judgment with the MySurgeryRisk Algorithm for Preoperative Risk Assessment: A Pilot Usability Study. *Surgery* 165 (5), pp. 1035–1045. ISSN: 1532-7361. DOI: 10.1016/j.surg.2019.01.002.

Calvert, J. S., Price, D. A., Chettipally, U. K., Barton, C. W., Feldman, M. D., Hoffman, J. L., Jay, M., & Das, R. (2016). A Computational Approach to Early Sepsis Detection. *Computers in Biology and Medicine* 74, pp. 69–73. ISSN: 00104825. DOI: 10.1016/j.compbiomed.2016.05.003.

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2020). *Shiny: Web Application Framework for R*.

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *New England Journal of Medicine* 378 (11), pp. 981–983. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMp1714229.

Che, Z., Cheng, Y., Zhai, S., Sun, Z., & Liu, Y. (2017). Boosting Deep Learning Risk Prediction with Generative Adversarial Networks for Electronic Health Records. In: *2017 IEEE International Conference on Data Mining (ICDM)*. New Orleans, LA: IEEE, pp. 787–792. ISBN: 978-1-5386-3835-4. DOI: 10.1109/ICDM.2017.93.

Coiera, E. W. (1996). Artificial Intelligence in Medicine: The Challenges Ahead. *Journal of the American Medical Informatics Association* 3 (6), pp. 363–366. ISSN: 1527-974X, 1067-5027. DOI: 10.1136/jamia.1996.97084510.

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. (2015). Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. *BMC Medicine* 13 (1), p. 1. ISSN: 1741-7015. DOI: 10.1186/s12916-014-0241-z.

Corradi, J. P., Thompson, S., Mather, J. F., Waszynski, C. M., & Dicks, R. S. (2018). Prediction of Incident Delirium Using a Random Forest Classifier. *Journal of Medical Systems* 42 (12). ISSN: 0148-5598, 1573-689X. DOI: 10.1007/s10916-018-1109-0.

Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* 16 (3), p. 38.

D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., & Kannel, W. B. (2008). General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. *Circulation* 117 (6), pp. 743–753. ISSN: 0009-7322, 1524-4539. DOI: 10.1161/CIRCULATIONAHA.107.699579.

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13 (3), p. 319. ISSN: 02767783. DOI: 10.2307/249008.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. *Management Science* 35 (8), pp. 982–1003. ISSN: 0025-1909, 1526-5501. DOI: 10.1287/mnsc.35.8.982.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* 44 (3), p. 837. ISSN: 0006341X. DOI: 10.2307/2531595.

Desautels, T., Calvert, J., Hoffman, J., Mao, Q., Jay, M., Fletcher, G., Barton, C., Chettipally, U., Kerem, Y., & Das, R. (2017). Using Transfer Learning for Improved Mortality Prediction in a Data-Scarce Hospital Setting. *Biomedical Informatics Insights* 9, p. 117822261771299. ISSN: 1178-2226, 1178-2226. DOI: 10.1177/1178222617712994.

Diefenbach, M. A., Weinstein, N. D., & O'Reilly, J. (1993). Scales for Assessing Perceptions of Health Hazard Susceptibility. *Health Education Research* 8 (2), pp. 181–192. ISSN: 0268-1153, 1465-3648. DOI: 10.1093/her/8.2.181.

Eeles, E. M. P., Hubbard, R. E., White, S. V., O'Mahony, M. S., Savva, G. M., & Bayer, A. J. (2010). Hospital Use, Institutionalisation and Mortality Associated with Delirium. *Age and Ageing* 39 (4), pp. 470–475. ISSN: 1468-2834 0002-0729. DOI: 10.1093/ageing/afq052.

Elfiky, A. A., Pany, M. J., Parikh, R. B., & Obermeyer, Z. (2018). Development and Application of a Machine Learning Approach to Assess Short-Term Mortality Risk Among Patients With Cancer Starting Chemotherapy. *JAMA Network Open* 1 (3), e180926. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2018.0926.

Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A Multiple Resampling Method for Learning from Imbalanced Data Sets. *Computational Intelligence* 20 (1), pp. 18–36. ISSN: 0824-7935, 1467-8640. DOI: 10.1111/j.0824-7935.2004.t01-1-00228.x.

Feinerer, I. & Hornik, K. (2020). *Tm: Text Mining Package.*

Fishbein, M & Ajzen, I (1975). *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*. MA: Addison-Wesley.

Fong, T. G., Davis, D., Growdon, M. E., Albuquerque, A., & Inouye, S. K. (2015). The Interface between Delirium and Dementia in Elderly Adults. *The Lancet Neurology* 14 (8), pp. 823–832. ISSN: 14744422. DOI: 10.1016/S1474-4422(15)00101-5.

Franke, R. H. & Kaul, J. D. (1978). The Hawthorne Experiments: First Statistical Interpretation. *American Sociological Review* 43 (5), p. 623. ISSN: 00031224. DOI: 10.2307/2094540.

Giannini, H. M., Ginestra, J. C., Chivers, C., Draugelis, M., Hanish, A., Schweickert, W. D., Fuchs, B. D., Meadows, L., Lynch, M., Donnelly, P. J., Pavan, K., Fishman, N. O., Hanson, C. W., & Umscheid, C. A. (2019). A Machine Learning Algorithm to Predict Severe Sepsis and Septic Shock: Development, Implementation, and Impact on Clinical Practice*. *Critical Care Medicine* 47 (11), pp. 1485–1492. ISSN: 0090-3493. DOI: 10.1097/CCM.0000000000003891.

Gijsberts, C. M., Groenewegen, K. A., Hoefer, I. E., Eijkemans, M. J. C., Asselbergs, F. W., Anderson, T. J., Britton, A. R., Dekker, J. M., Engström, G., Evans, G. W., de Graaf, J., Grobbee, D. E., Hedblad, B., Holewijn, S., Ikeda, A., Kitagawa, K., Kitamura, A., de Kleijn, D. P. V., Lonn, E. M., Lorenz, M. W., Mathiesen, E. B., Nijpels, G., Okazaki, S., O'Leary, D. H., Pasterkamp, G., Peters, S. A. E., Polak, J. F., Price, J. F., Robertson, C., Rembold, C. M., Rosvall, M., Rundek, T., Salonen, J. T., Sitzer, M., Stehouwer, C. D. A., Bots, M. L., & den Ruijter, H. M. (2015). Race/Ethnic Differences in the Associations of the Framingham Risk Factors with Carotid IMT and Cardiovascular Events. *PLOS ONE* 10 (7). Ed. by C. Apetrei, e0132321. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0132321.

Ginestra, J. C., Giannini, H. M., Schweickert, W. D., Meadows, L., Lynch, M. J., Pavan, K., Chivers, C. J., Draugelis, M., Donnelly, P. J., Fuchs, B. D., & Umscheid, C. A. (2019). Clinician Perception of a Machine Learning–Based Early Warning System Designed to Predict Severe Sepsis and Septic Shock. *Critical Care Medicine* 47 (11), pp. 1477–1484. ISSN: 0090-3493. DOI: 10.1097/CCM.0000000000003803.

Girard, T. D., Jackson, J. C., Pandharipande, P. P., Pun, B. T., Thompson, J. L., Shintani, A. K., Gordon, S. M., Canonico, A. E., Dittus, R. S., Bernard, G. R., & Wesley Ely, E. (2010). Delirium as a Predictor of Long-Term Cognitive Impairment in Survivors of Critical Illness. *Critical Care Medicine* 38 (7), pp. 1513–1520. ISSN: 0090-3493. DOI: 10.1097/CCM.0b013e3181e47be1.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators. *Journal of the American Medical Informatics Association* 19 (1), pp. 121–127. ISSN: 1067-5027, 1527-974X. DOI: 10.1136/amiajnl-2011-000089.

Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. P. A. (2017). Opportunities and Challenges in Developing Risk Prediction Models with Electronic Health Records Data: A Systematic Review. *Journal of the American Medical Informatics Association* 24 (1), pp. 198–208. ISSN: 1067-5027, 1527-974X. DOI: 10.1093/jamia/ocw042.

Halladay, C. W., Sillner, A. Y., & Rudolph, J. L. (2018). Performance of Electronic Prediction Rules for Prevalent Delirium at Hospital Admission. *JAMA Network Open* 1 (4), e181405. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2018.1405.

Hand, D. J. (2012). Assessing the Performance of Classification Methods. *International Statistical Review* 80 (3), pp. 400–414. ISSN: 03067734. DOI: 10.1111/j.1751-5823.2012.00183.x.

Hao, S., Wang, Y., Jin, B., Shin, A. Y., Zhu, C., Huang, M., Zheng, L., Luo, J., Hu, Z., Fu, C., Dai, D., Wang, Y., Culver, D. S., Alfreds, S. T., Rogow, T., Stearns, F., Sylvester, K. G., Widen, E., & Ling, X. B. (2015). Development, Validation and Deployment of a Real Time 30 Day Hospital Readmission Risk Assessment Tool in the Maine Healthcare Information Exchange. *PLOS ONE* 10 (10). Ed. by J. I. Salluh, e0140271. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0140271.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Second. New York, NY: Springer.

He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The Practical Implementation of Artificial Intelligence Technologies in Medicine. *Nature Medicine* 25 (1), pp. 30–36. ISSN: 1078-8956. DOI: 10.1038/s41591-018-0307-0.

Holden, R. J. & Karsh, B.-T. (2010). The Technology Acceptance Model: Its Past and Its Future in Health Care. *Journal of Biomedical Informatics* 43 (1), pp. 159–172. ISSN: 15320464. DOI: 10.1016/j.jbi.2009.07.002.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. third edition. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley. ISBN: 978-0-470-58247-3 978-1-118-54838-7.

Hshieh, T. T., Yue, J., Oh, E., Puelle, M., Dowal, S., Travison, T., & Inouye, S. K. (2015). Effectiveness of Multicomponent Nonpharmacological Delirium Interventions: A Meta-Analysis. *JAMA Internal Medicine* 175 (4), p. 512. ISSN: 2168-6106. DOI: 10.1001/jamainternmed.2014.7779.

Inouye, S. K. (2006). Delirium in Older Persons. *New England Journal of Medicine* 354 (11), pp. 1157–1165. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMra052321.

Inouye, S. K., Bogardus Jr, S. T., Charpentier, P. A., Leo-Summers, L., Acampora, D., Holford, T. R., & Cooney Jr, L. M. (1999). A Multicomponent Intervention to Prevent Delirium in Hospitalized Older Patients. *New England Journal of Medicine* 340 (9), pp. 669–676.

Inouye, S. K., Christopher, M., Dyck, H., Cathy, M., Alessi, A., Balkin, M. S., Alan, M., Siegal, P., & Horwitz, R. I. (1990). Clarifying Confusion: The Confusion Assessment Method. A New Method for Detection of Delirium. *Annals of Internal Medicine* 113 (12), pp. 941–948.

Inouye, S. K., Westendorp, R. G., & Saczynski, J. S. (2014). Delirium in Elderly People. *The Lancet* 383 (9920), pp. 911–922. ISSN: 01406736. DOI: 10.1016/S0140-6736(13)60688-1.

Islam, M., Hasan, M., Wang, X., Germack, H., & Noor-E-Alam, M. (2018). A Systematic Review on Healthcare Analytics: Application and Theoretical Perspective of Data Mining. *Healthcare* 6 (2), p. 54. ISSN: 2227-9032. DOI: 10.3390/healthcare6020054.

Janssen, F. & Kunst, A. E. (2004). ICD Coding Changes and Discontinuities in Trends in Cause- Specific Mortality in Six European Countries, 1950–99. *Bulletin of the World Health Organization*, p. 12.

Jauk, S. (2021). Reply to Rousseau and Tierney. *Journal of the American Medical Informatics Association* 28 (3), pp. 666–667. ISSN: 1527-974X. DOI: 10.1093/jamia/ocaa286.

Jauk, S., Kramer, D., Avian, A., Berghold, A., Leodolter, W., & Schulz, S. (2021). Technology Acceptance of a Machine Learning Algorithm Predicting Delirium in a Clinical Setting: A Mixed-Methods Study. *Journal of Medical Systems* 45 (48). ISSN: 0148-5598, 1573-689X. DOI: 10.1007/s10916-021-01727-6.

Jauk, S., Kramer, D., Großauer, B., Rienmüller, S., Avian, A., Berghold, A., Leodolter, W., & Schulz, S. (2020). Risk Prediction of Delirium in Hospitalized Patients Using Machine Learning: An Implementation and Prospective Evaluation Study. *Journal of the American Medical Informatics Association* 27 (9), pp. 1383–1392. DOI: 10.1093/jamia/ocaa113.

Jauk, S., Kramer, D., Quehenberger, F., Veeranki Sai Pavan, K., Hayn, D., Schreier, G., & Leodolter, W. (2019). Information Adapted Machine Learning Models for Prediction

in Clinical Workflow. *Studies in Health Technology and Informatics* 260, pp. 65–72. ISSN: 0926-9630. DOI: 10.3233/978-1-61499-971-3-65.

Jenkins, D. A., Sperrin, M., Martin, G. P., & Peek, N. (2018). Dynamic Models to Predict Health Outcomes: Current Status and Methodological Challenges. *Diagnostic and Prognostic Research* 2 (1), p. 23. ISSN: 2397-7523. DOI: 10.1186/s41512-018-0045-2.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial Intelligence in Healthcare: Past, Present and Future. *Stroke and Vascular Neurology* 2 (4), pp. 230–243. ISSN: 2059-8688, 2059-8696. DOI: 10.1136/svn-2017-000101.

Kim, M. Y., Park, U. J., Kim, H. T., & Cho, W. H. (2016). DELirium Prediction Based on Hospital Information (Delphi) in General Surgery Patients. *Medicine* 95 (12), e3072. ISSN: 1536-5964. DOI: 10.1097/MD.0000000000003072.

King, W. R. & He, J. (2006). A Meta-Analysis of the Technology Acceptance Model. *Information & Management* 43 (6), pp. 740–755. ISSN: 03787206. DOI: 10.1016/j.im.2006.05.003.

Kramer, D., Veeranki, S., Hayn, D., Quehenberger, F., Leodolter, W., Jagsch, C., & Schreier, G. (2017). Development and Validation of a Multivariable Prediction Model for the Occurrence of Delirium in Hospitalized Gerontopsychiatry and Internal Medicine Patients. *Studies in Health Technology and Informatics* 236, pp. 32–39.

Kuhn, M. (2017). *Caret: Classification and Regression Training. R Package Version 6.0-78.*

Kuhn, M. & Johnson, K. (2013). *Applied Predictive Modeling*. New York: Springer-Verlag. ISBN: 978-1-4614-6848-6.

Lange, P. W., Lamanna, M., Watson, R., & Maier, A. B. (2019). Undiagnosed Delirium Is Frequent and Difficult to Predict: Results from a Prevalence Survey of a Tertiary Hospital. *Journal of Clinical Nursing*, jocn.14833. ISSN: 0962-1067, 1365-2702. DOI: 10.1111/jocn.14833.

Laupacis, A., Wells, G., Richardson, S., & Tugwell, P. (1994). Users' Guides to the Medical Literature–V. How to Use an Article about Prognosis. *JAMA* 272 (3), p. 234.

Lee, C. H. & Yoon, H.-J. (2017). Medical Big Data: Promise and Challenges. *Kidney Research and Clinical Practice* 36 (1), pp. 3–11. ISSN: 2211-9140. DOI: 10.23876/j.krcp.2017.36.1.3.

Lee, J. H., Jang, M. K., Lee, J. Y., Kim, S. M., Kim, K. H., Park, J. Y., Lee, J. H., Kim, H. Y., & Yoo, J. Y. (2005). Clinical Predictors for Delirium Tremens in Alcohol Dependence. *Journal of Gastroenterology and Hepatology* 20, pp. 1833–1837.

Lee, T. C., Shah, N. U., Haack, A., & Baxter, S. L. (2020). Clinical Implementation of Predictive Models Embedded within Electronic Health Record Systems: A Systematic Review. *Informatics* 7 (3), p. 25. ISSN: 2227-9709. DOI: 10.3390/informatics7030025.

Lenert, M. C., Matheny, M. E., & Walsh, C. G. (2019). Prognostic Models Will Be Victims of Their Own Success, Unless... *Journal of the American Medical Informatics Association* 26 (12), pp. 1645–1650. ISSN: 1527-974X. DOI: 10.1093/jamia/ocz145.

Leslie, D. L. (2008). One-Year Health Care Costs Associated With Delirium in the Elderly Population. *Archives of Internal Medicine* 168 (1), p. 27. ISSN: 0003-9926. DOI: 10.1001/archinternmed.2007.4.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., Clarke, M., Devereaux, P., Kleijnen, J., & Moher, D. (2009). The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *Journal of Clinical Epidemiology* 62 (10), e1–e34. ISSN: 08954356. DOI: 10.1016/j.jclinepi.2009.06.006.

Lin, S.-M., Liu, C.-Y., Wang, C.-H., Lin, H.-C., Huang, C.-D., Huang, P.-Y., Fang, Y.-F., Shieh, M.-H., & Kuo, H.-P. (2004). The Impact of Delirium on the Survival of Mechanically Ventilated Patients*: *Critical Care Medicine* 32 (11), pp. 2254–2259. ISSN: 0090-3493. DOI: 10.1097/01.CCM.0000145587.16421.BB.

Lindroth, H., Bratzke, L., Purvis, S., Brown, R., Coburn, M., Mrkobrada, M., Chan, M. T. V., Davis, D. H. J., Pandharipande, P., Carlsson, C. M., & Sanders, R. D. (2018). Systematic Review of Prediction Models for Delirium in the Older Adult Inpatient. *BMJ Open* 8 (4), e019223. ISSN: 2044-6055, 2044-6055. DOI: 10.1136/bmjopen-2017-019223.

Lipton, Z. C., Kale, D. C., Elkan, C., & Wetzel, R. C. (2016). Learning to Diagnose with LSTM Recurrent Neural Networks. In: *4th International Conference on Learning Representations (ICLR 2016)*. Ed. by Y. Bengio & Y. LeCun.

Liu, S., Du, H., & Feng, M. (2020). "Robust Predictive Models in Clinical Data—Random Forest and Support Vector Machines". In: *Leveraging Data Science for Global Health*. Ed. by L. A. Celi, M. S. Majumder, P. Ordóñez, J. S. Osorio, K. E. Paik, & M. Somai. Cham: Springer International Publishing, pp. 219–228. ISBN: 978-3-030-47994-7. DOI: 10.1007/978-3-030-47994-7_13.

Lyell, D. & Coiera, E. (2017). Automation Bias and Verification Complexity: A Systematic Review. *Journal of the American Medical Informatics Association* 24 (2), pp. 423–431. ISSN: 1067-5027, 1527-974X. DOI: 10.1093/jamia/ocw105.

Magrabi, F., Ammenwerth, E., McNair, J., De Keizer, N., Hyppönen, H., Nykänen, P., Rigby, M., Scott, P., Vehko, T., Wong, Z., & Georgiou, A. (2019). Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications: A Position Paper from the IMIA Technology Assessment & Quality Development in Health Informatics Working Group and the EFMI Working Group for Assessment of

Health Information Systems. *Yearbook of Medical Informatics*. ISSN: 0943-4747, 2364-0502. DOI: 10.1055/s-0039-1677903.

Mainerova, B., Prasko, J., Latalova, K., Axmann, K., Cerna, M., Horacek, R., & Bradacova, R. (2015). Alcohol Withdrawal Delirium - Diagnosis, Course and Treatment. *Biomedical Papers* 159 (1), pp. 044–052. ISSN: 12138118, 18047521. DOI: 10.5507/bp.2013.089.

McCusker, J., Cole, M. G., Dendukuri, N., & Belzile, E. (2003). Does Delirium Increase Hospital Stay? *Journal of the American Geriatrics Society* 51 (11), pp. 1539–1546. DOI: 10.1046/j.1532-5415.2003.51509.x.

McCusker, J., Cole, M., Abrahamowicz, M., Han, L., Podoba, J. E., & Ramman-Haddad, L. (2001). Environmental Risk Factors for Delirium in Hospitalized Older People. *Journal of the American Geriatrics Society* 49 (10), pp. 1327–1334. ISSN: 0002-8614, 1532-5415. DOI: 10.1046/j.1532-5415.2001.49260.x.

Michie, S. & Marteau, T. (1999). Non-Response Bias in Prospective Studies of Patients and Health Care Professionals. *International Journal of Social Research Methodology* 2 (3), pp. 203–212. ISSN: 1364-5579, 1464-5300. DOI: 10.1080/136455799295014.

Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., Macaskill, P., Steyerberg, E. W., Vickers, A. J., Ransohoff, D. F., & Collins, G. S. (2015). Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Annals of Internal Medicine* 162 (1), W1. ISSN: 0003-4819. DOI: 10.7326/M14-0698.

Müller-Riemenschneider, F. (2010). Barriers to Routine Risk-Score Use for Healthy Primary Care Patients: Survey and Qualitative Study. *Archives of Internal Medicine* 170 (8), p. 719. ISSN: 0003-9926. DOI: 10.1001/archinternmed.2010.66.

Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H., Topol, E. J., Ioannidis, J. P. A., Collins, G. S., & Maruthappu, M. (2020). Artificial Intelligence versus Clinicians: Systematic Review of Design, Reporting Standards, and Claims of Deep Learning Studies. *BMJ*, p. m689. ISSN: 1756-1833. DOI: 10.1136/bmj.m689.

DELIRIUM: Diagnosis, Prevention and Management (2010). Tech. rep. London: National Clinical Guideline Centre.

Newman, M. W., O'Dwyer, L. C., & Rosenthal, L. (2015). Predicting Delirium: A Review of Risk-Stratification Models. *General Hospital Psychiatry* 37 (5), pp. 408–413. ISSN: 01638343. DOI: 10.1016/j.genhosppsych.2015.05.003.

O'Malley, K. J., Cook, K. F., Price, M. D., Wildes, K. R., Hurdle, J. F., & Ashton, C. M. (2005). Measuring Diagnoses: ICD Code Accuracy. *Health Services Research* 40 (5p2), pp. 1620–1639. ISSN: 0017-9124, 1475-6773. DOI: 10.1111/j.1475-6773.2005.00444.x.

Parikh, R. B., Kakad, M., & Bates, D. W. (2016). Integrating Predictive Analytics Into High-Value Care: The Dawn of Precision Delivery. *JAMA* 315 (7), p. 651. ISSN: 0098-7484. DOI: 10.1001/jama.2015.19417.

Peek, N., Combi, C., Marin, R., & Bellazzi, R. (2015). Thirty Years of Artificial Intelligence in Medicine (AIME) Conferences: A Review of Research Themes. *Artificial Intelligence in Medicine* 65 (1), pp. 61–73. ISSN: 09333657. DOI: 10.1016/j.artmed.2015.07.003.

Perry, W. M., Hossain, R., & Taylor, R. A. (2018). Assessment of the Feasibility of Automated, Real-Time Clinical Decision Support in the Emergency Department Using Electronic Health Record Data. *BMC Emergency Medicine* 18 (1). ISSN: 1471-227X. DOI: 10.1186/s12873-018-0170-9.

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. *New England Journal of Medicine* 380 (14), pp. 1347–1358. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMra1814259.

Reddy, C. K. & Li, Y. (2015). A Review of Clinical Prediction Models. *Healthcare data analytics* 36, pp. 343–378.

Rehm, J., Allamani, A., Elekes, Z., Jakubczyk, A., Manthey, J., Probst, C., Struzzo, P., Della Vedova, R., Gual, A., & Wojnar, M. (2015). Alcohol Dependence and Treatment Utilization in Europe – a Representative Cross-Sectional Study in Primary Care. *BMC Family Practice* 16 (1), p. 90. ISSN: 1471-2296. DOI: 10.1186/s12875-015-0308-8.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco California USA: ACM, pp. 1135–1144. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778.

Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The Future of Digital Health with Federated Learning. *npj Digital Medicine* 3 (1), p. 119. ISSN: 2398-6352. DOI: 10.1038/s41746-020-00323-1.

Rizopoulos, D. (2018). *Latent Trait Models under IRT*.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). *pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves*.

Rousseau, J. F. & Tierney, W. M. (2021). Letter to the Editor in Response to "Risk Prediction of Delirium in Hospitalized Patients Using Machine Learning: An Implementation and Prospective Evaluation Study". *Journal of the American Medical Informatics Association* 28 (3), pp. 664–665. ISSN: 1527-974X. DOI: 10.1093/jamia/ocaa285.

Rudolph, J. L., Doherty, K., Kelly, B., Driver, J. A., & Archambault, E. (2016). Validation of a Delirium Risk Assessment Using Electronic Medical Record Information. *Journal of the American Medical Directors Association* 17 (3), pp. 244–248. ISSN: 15258610. DOI: 10.1016/j.jamda.2015.10.020.

Ryan, D. J., O'Regan, N. A., Caoimh, R. Ó., Clare, J., O'Connor, M., Leonard, M., McFarland, J., Tighe, S., O'Sullivan, K., Trzepacz, P. T., Meagher, D., & Timmons, S. (2013). Delirium in an Adult Acute Hospital Population: Predictors, Prevalence and Detection. *BMJ Open* 3 (1), e001772. ISSN: 2044-6055, 2044-6055. DOI: 10.1136/bmjopen-2012-001772.

Sabetta, L. (2019). "Self-Defeating Prophecies: When Sociology Really Matters". In: *Anticipation, Agency and Complexity*. Ed. by R. Poli & M. Valerio. Cham: Springer International Publishing, pp. 51–59. ISBN: 978-3-030-03623-2. DOI: 10.1007/978-3-030-03623-2_4.

Saito, T. & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE* 10 (3). Ed. by G. Brock, e0118432. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0118432.

Schmitt, E. M., Gallagher, J., Albuquerque, A., Tabloski, P., Lee, H. J., Gleason, L., Weiner, L. S., Marcantonio, E. R., Jones, R. N., Inouye, S. K., & Schulman-Green, D. (2019). Perspectives on the Delirium Experience and Its Burden: Common Themes Among Older Patients, Their Family Caregivers, and Nurses. *The Gerontologist* 59 (2), pp. 327–337. ISSN: 0016-9013, 1758-5341. DOI: 10.1093/geront/gnx153.

Schuckit, M. A. (2014). Recognition and Management of Withdrawal Delirium (Delirium Tremens). *New England Journal of Medicine* 371 (22). Ed. by D. L. Longo, pp. 2109–2113. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMra1407298.

Schuurmans, M., Shortridge-Baggett, L., & Duursma, S. (2003). The Delirium Observation Screening Scale: A Screening Instrument for Delirium. *Research and Theory for Nursing Practice* 17 (1), pp. 31–50.

Shen, J., Zhang, C. J. P., Jiang, B., Chen, J., Song, J., Liu, Z., He, Z., Wong, S. Y., Fang, P.-H., & Ming, W.-K. (2019). Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review. *JMIR Medical Informatics* 7 (3), e10010. ISSN: 2291-9694. DOI: 10.2196/10010.

Shiffrin, R. M. (2016). Drawing Causal Inference from Big Data. *Proceedings of the National Academy of Sciences* 113 (27), pp. 7308–7309. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1608845113.

Shimabukuro, D. W., Barton, C. W., Feldman, M. D., Mataraso, S. J., & Das, R. (2017). Effect of a Machine Learning-Based Severe Sepsis Prediction Algorithm on Patient

Survival and Hospital Length of Stay: A Randomised Clinical Trial. *BMJ Open Respiratory Research* 4 (1), e000234. ISSN: 2052-4439. DOI: 10.1136/bmjresp-2017-000234.

Siddiqi, N., Harrison, J. K., Clegg, A., Teale, E. A., Young, J., Taylor, J., & Simpkins, S. A. (2016). Interventions for Preventing Delirium in Hospitalised Non-ICU Patients. *Cochrane Database of Systematic Reviews*. Ed. by Cochrane Dementia and Cognitive Improvement Group. ISSN: 14651858. DOI: 10.1002/14651858.CD005563.pub3.

Siddiqi, N., House, A. O., & Holmes, J. D. (2006). Occurrence and Outcome of Delirium in Medical In-Patients: A Systematic Literature Review. *Age and Ageing* 35 (4), pp. 350–364. ISSN: 1468-2834, 0002-0729. DOI: 10.1093/ageing/afl005.

Spiegelhalter, D. (2020). Should We Trust Algorithms? *Harvard Data Science Review*. DOI: 10.1162/99608f92.cb91a35a.

Stausberg, J., Lehmann, N., Kaczmarek, D., & Stein, M. (2008). Reliability of Diagnoses Coding with ICD-10. *International Journal of Medical Informatics* 77 (1), pp. 50–57. ISSN: 13865056. DOI: 10.1016/j.ijmedinf.2006.11.005.

Steyerberg, E. W. (2019). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Statistics for Biology and Health. Cham: Springer International Publishing. ISBN: 978-3-030-16398-3 978-3-030-16399-0. DOI: 10.1007/978-3-030-16399-0.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. *Epidemiology* 21 (1), pp. 128–138. ISSN: 1044-3983. DOI: 10.1097/EDE.0b013e3181c30fb2.

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics* 8 (1). ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-25.

Tavakol, M. & Dennick, R. (2011). Making Sense of Cronbach's Alpha. *International Journal of Medical Education* 2, pp. 53–55. ISSN: 20426372. DOI: 10.5116/ijme.4dfb.8dfd.

Topol, E. J. (2019). High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nature Medicine* 25 (1), pp. 44–56. ISSN: 1078-8956, 1546-170X. DOI: 10.1038/s41591-018-0300-7.

Turing, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind* LIX (236), pp. 433–460. ISSN: 1460-2113, 0026-4423. DOI: 10.1093/mind/LIX.236.433.

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles Heel of Predictive Analytics. *BMC Medicine* 17 (1), p. 230. ISSN: 1741-7015. DOI: 10.1186/s12916-019-1466-7.

Van Calster, B., Nieboer, D., Vergouwe, Y., Cock, B. D., Pencina, M. J., & Steyerberg, E. W. (2016). A Calibration Hierarchy for Risk Models Was Defined: From Utopia to Empirical Data. *Journal of Clinical Epidemiology* 74, pp. 167–176. ISSN: 0895-4356, 1878-5921. DOI: 10.1016/j.jclinepi.2015.12.005.

VanHouten, J. P., Starmer, J. M., Lorenzi, N. M., Maron, D. J., & Lasko, T. A. (2014). Machine Learning for Risk Prediction of Acute Coronary Syndrome. *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2014, pp. 1940–1949. ISSN: 1942-597X.

Varonen, H., Kortteisto, T., Kaila, M., & for the EBMeDS Study Group (2008). What May Help or Hinder the Implementation of Computerized Decision Support Systems (CDSSs): A Focus Group Study with Physicians. *Family Practice* 25 (3), pp. 162–167. ISSN: 0263-2136, 1460-2229. DOI: 10.1093/fampra/cmn020.

Veeranki, S., Kramer, D, Hayn, D, Jauk, S, Eggerth, Quehenberger, F, Leodolter, W, & Schreier, G (2019). Is Regular Re-Training of a Predictive Delirium Model Necessary After Deployment in Routine Care? *Studies in Health Technology and Informatics* 260, pp. 186–191. ISSN: 1879-8365 (electronic).

Veeranki, S., Hayn, D., Eggerth, A., Jauk, S., Kramer, D., Leodolter, W., & Schreier, G. (2018). On the Representation of Machine Learning Results for Delirium Prediction in a Hospital Information System in Routine Care. *Studies in Health Technology and Informatics* 251, pp. 97–100. ISSN: 0926-9630. DOI: 10.3233/978-1-61499-880-8-97.

Venkatesh, V. & Davis, F. D. (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science* 46 (2), pp. 186–204.

Vollmer, S., Mateen, B. A., Bohner, G., Király, F. J., Ghani, R., Jonsson, P., Cumbers, S., Jonas, A., McAllister, K. S. L., Myles, P., Grainger, D., Birse, M., Branson, R., Moons, K. G. M., Collins, G. S., Ioannidis, J. P. A., Holmes, C., & Hemingway, H. (2020). Machine Learning and Artificial Intelligence Research for Patient Benefit: 20 Critical Questions on Transparency, Replicability, Ethics, and Effectiveness. *BMJ*, p. l6927. ISSN: 1756-1833. DOI: 10.1136/bmj.l6927.

Wang, L., Shaw, P. A., Mathelier, H. M., Kimmel, S. E., & French, B. (2016). Evaluating Risk-Prediction Models Using Data from Electronic Health Records. *The Annals of Applied Statistics* 10 (1), pp. 286–304. ISSN: 1932-6157. DOI: 10.1214/15-AOAS891.

Waring, J., Lindvall, C., & Umeton, R. (2020). Automated Machine Learning: Review of the State-of-the-Art and Opportunities for Healthcare. *Artificial Intelligence in Medicine* 104, p. 101822. ISSN: 09333657. DOI: 10.1016/j.artmed.2020.101822.

Watson, J., Hutyra, C. A., Clancy, S. M., Chandiramani, A., Bedoya, A., Ilangovan, K., Nderitu, N., & Poon, E. G. (2020). Overcoming Barriers to the Adoption and Implementation of Predictive Modeling and Machine Learning in Clinical Care: What

Can We Learn from US Academic Medical Centers? *JAMIA Open* 3 (2), pp. 167–172. ISSN: 2574-2531. DOI: 10.1093/jamiaopen/ooz046.

Watt, J., Tricco, A. C., Talbot-Hamon, C., Pham, B., Rios, P., Grudniewicz, A., Wong, C., Sinclair, D., & Straus, S. E. (2018). Identifying Older Adults at Risk of Delirium Following Elective Surgery: A Systematic Review and Meta-Analysis. *Journal of General Internal Medicine* 33 (4), pp. 500–509. ISSN: 0884-8734, 1525-1497. DOI: 10.1007/s11606-017-4204-x.

Weinrebe, W., Johannsdottir, E., Karaman, M., & Füsgen, I. (2016). What Does Delirium Cost?: An Economic Evaluation of Hyperactive Delirium. *Zeitschrift für Gerontologie und Geriatrie* 49 (1), pp. 52–58. ISSN: 0948-6704, 1435-1269. DOI: 10.1007/s00391-015-0871-6.

Weiskopf, N. G. & Weng, C. (2013). Methods and Dimensions of Electronic Health Record Data Quality Assessment: Enabling Reuse for Clinical Research. *Journal of the American Medical Informatics Association: JAMIA* 20 (1), pp. 144–151. ISSN: 1527-974X. DOI: 10.1136/amiajnl-2011-000681.

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can Machine-Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data? *PLOS ONE* 12 (4). Ed. by B. Liu, e0174944. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0174944.

Weng, W.-H. (2020). "Machine Learning for Clinical Predictive Analytics". In: *Leveraging Data Science for Global Health*. Ed. by L. A. Celi, M. S. Majumder, P. Ordóñez, J. S. Osorio, K. E. Paik, & M. Somai. Cham: Springer International Publishing, pp. 199–217. ISBN: 978-3-030-47994-7. DOI: 10.1007/978-3-030-47994-7_12.

Witlox, J., Eurelings, L. S. M., de Jonghe, J. F. M., Kalisvaart, K. J., Eikelenboom, P., & van Gool, W. A. (2010). Delirium in Elderly Patients and the Risk of Postdischarge Mortality, Institutionalization, and Dementia: A Meta-Analysis. *JAMA* 304 (4), pp. 443–451. ISSN: 0098-7484. DOI: 10.1001/jama.2010.1013.

Wong, A., Young, A. T., Liang, A. S., Gonzales, R., Douglas, V. C., & Hadley, D. (2018). Development and Validation of an Electronic Health Record–Based Machine Learning Model to Estimate Delirium Risk in Newly Hospitalized Patients Without Known Cognitive Impairment. *JAMA Network Open* 1 (4), e181018. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2018.1018.

Wyatt, J & Spiegelhalter, D (1990). Evaluating Medical Expert Systems: What to Test and How? *Med Inform (Lond).* 15 (3), pp. 205–17. DOI: 10.3109/14639239009025268.

Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and Challenges in Developing Deep Learning Models Using Electronic Health Records Data: A Systematic Review. *Journal*

of the American Medical Informatics Association 25 (10), pp. 1419–1428. ISSN: 1067-5027, 1527-974X. DOI: 10.1093/jamia/ocy068.

Yang, F. M., Marcantonio, E. R., Inouye, S. K., Kiely, D. K., Rudolph, J. L., Fearing, M. A., & Jones, R. N. (2009). Phenomenological Subtypes of Delirium in Older Persons: Patterns, Prevalence, and Prognosis. *Psychosomatics* 50 (3), pp. 248–254. ISSN: 00333182. DOI: 10.1176/appi.psy.50.3.248.

Young, J., Leentjens, A. F., George, J., Olofsson, B., & Gustafson, Y. (2008). Systematic Approaches to the Prevention and Management of Patiens with Delirium. *Journal of Psychosomatic Research* 65 (3), pp. 267–272. ISSN: 00223999. DOI: 10.1016/j.jpsychores.2008.05.022.

de Rooij, S. E., Schuurmans, M. J., van der Mast, R. C., & Levi, M. (2005). Clinical Subtypes of Delirium and Their Relevance for Daily Clinical Practice: A Systematic Review. *International Journal of Geriatric Psychiatry* 20 (7), pp. 609–615. ISSN: 0885-6230, 1099-1166. DOI: 10.1002/gps.1343.

de Wit, H. A. J. M., Winkens, B., Mestres Gonzalvo, C., Hurkens, K. P. G. M., Mulder, W. J., Janknegt, R., Verhey, F. R., van der Kuy, P.-H. M., & Schols, J. M. G. A. (2016). The Development of an Automated Ward Independent Delirium Risk Prediction Model. *International Journal of Clinical Pharmacy* 38 (4), pp. 915–923. ISSN: 2210-7703, 2210-7711. DOI: 10.1007/s11096-016-0312-7.

van Meenen, L. C. C., van Meenen, D. M. P., de Rooij, S. E., & ter Riet, G. (2014). Risk Prediction Models for Postoperative Delirium: A Systematic Review and Meta-Analysis. *Journal of the American Geriatrics Society* 62 (12), pp. 2383–2390. ISSN: 00028614. DOI: 10.1111/jgs.13138.

A

FIGURES

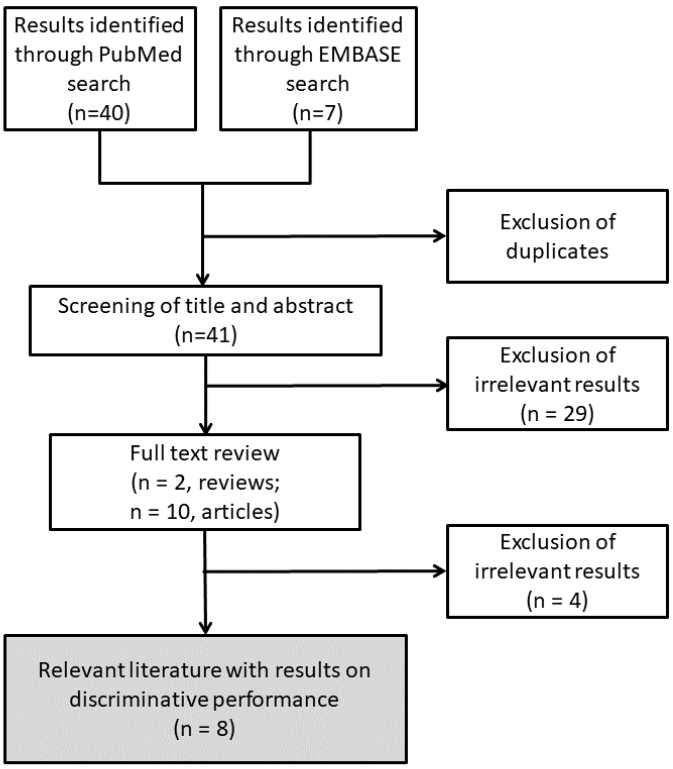


Figure A.1: PRISMA flow diagram (Liberati et al., 2009) with corresponding research items for the systematic review. The systematic review resulted in eight research items.

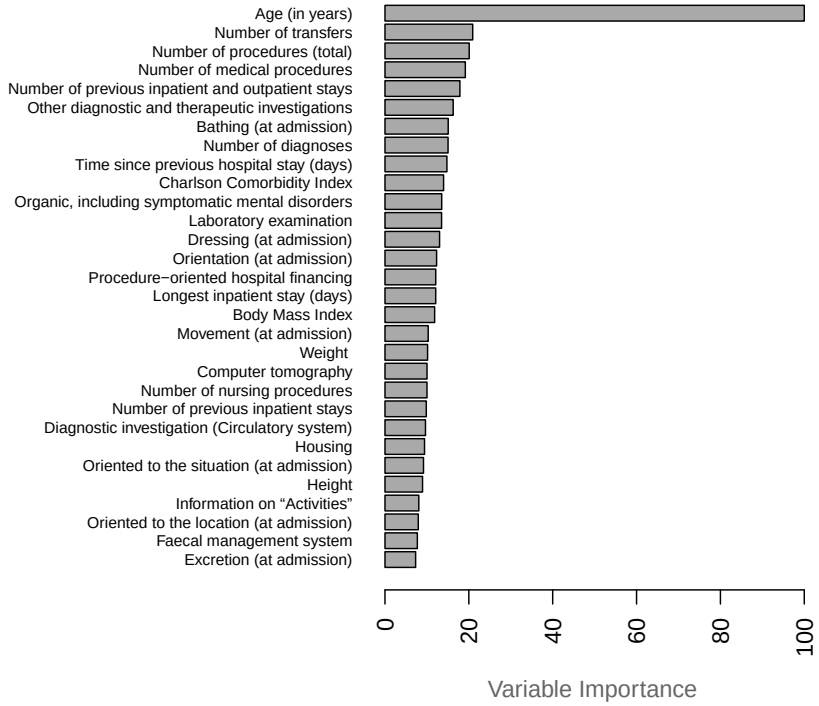


Figure A.2: Variable importance plot illustrating the 30 most important predictors in the random forest model predicting delirium coded with F05.

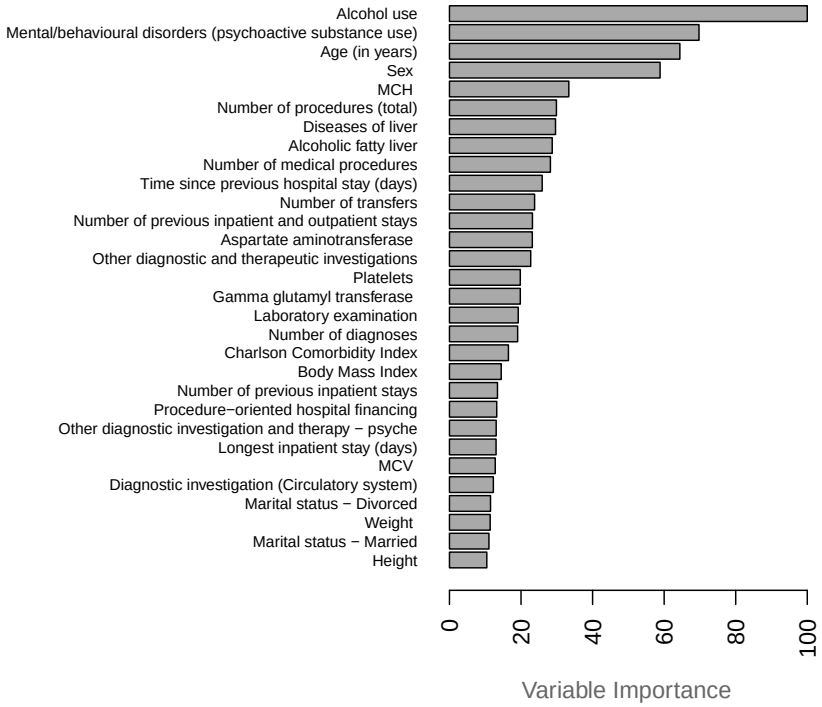


Figure A.3: Variable importance plot illustrating the 30 most important predictors in the random forest model predicting delirium coded with F10.4.

Feedback-Protokoll Delirprognose

Station:

		Ich schätze das Risiko eines Delirs ...					Ist eine Demenz bei dem Patienten bekannt?		Gibt es begründete Anzeichen für eine akute Veränderung des mentalen Status des Patienten?		Fluktuiert das (veränderte) Verhalten während des Tages, d.h. hatte es die Tendenz aufzutreten und wieder zu verschwinden, oder wurde es stärker und schwächer?		Hatte der Patient Schwierigkeiten, seine Aufmerksamkeit zu fokussieren, z.B. war er leicht ablenkbar oder hatte er Schwierigkeiten, dem Gespräch zu folgen?		War der Gedankenablauf des Patienten desorganisiert oder zusammenhangslos, wie Gefasel oder belanglose Konversation, unklarer oder unlogischer Gedankenfluss, oder unerwartete Gedankensprünge?		Wie würden Sie die Bewusstseinslage des Patienten allgemein beschreiben? Wach - alert (normal)?		Kommentare bei Entlassung:
Patienten-Nr	Datum	sehr niedrig	eher niedrig	mittel	eher hoch	sehr hoch	nein	ja	nein	ja	nein	ja	nein	ja	nein	ja	nein	ja	

Figure A.4: Documentation protocol used for the assessment of the delirium risk rated by clinical experts. The protocol includes one item for the risk rating of delirium, one item to assess an existing dementia, and five items of the CAM (Inouye et al., 1990).

EVALUIERUNG – PILOTPROJEKT DELIR-PROGNOSE

Um die Qualität der Delir-Prognose zu verbessern, ist eine Evaluierung der Anwendung vorgesehen. Mithilfe der folgenden Fragen werden Ihre Erfahrungen mit dem Delir-Prognose Tool erfasst. Alle Ihre Antworten werden vertrauensvoll behandelt und sind nicht auf Ihre Person zurückzuführen.

Um die verschiedenen Teile der Evaluierung zusammenführen zu können, wird für Sie ein persönlicher **Identifikationscode** erstellt, welcher sich wie folgt zusammenstellt:

1 2	3 4	5 6	
Erster und letzter Buchstabe des Vornamens Ihres Vaters (z.B. **K**ar**l** → KL)	Erster und letzter Buchstabe des Vornamens Ihrer Großmutter mütterlicherseits (z.B. **J**uli**a** → JA)	Letzten beiden Ziffern des Geburtsjahrs Ihrer Mutter (z.B. 04.07.19**48** → 48)	*Falls Sie einzelne Angaben nicht machen können, dann tragen Sie bitte 99 ein.*

Berufsgruppe:	☐ Arzt/Ärztin in Ausbildung ☐ Facharzt ☐ Pflegepersonal
Geschlecht:	☐ männlich ☐ weiblich
Alter:	

Bitte markieren Sie für folgende Aussagen jene Kategorie, die am besten auf Sie zutrifft.

	Trifft nicht zu	**Trifft wenig zu**	**Teils, teils**	**Trifft eher zu**	**Trifft sehr zu**
Der Zweck des Tools war klar und verständlich.	☐	☐	☐	☐	☐
Das Tool ist nützlich für meine Arbeit.	☐	☐	☐	☐	☐
Ich war zu Beginn des Projekts ausreichend vorbereitet um das Tool verwenden zu können.	☐	☐	☐	☐	☐
Ich glaube, dass das Tool eine sinnvolle Unterstützung in der Delir-Prophylaxe ist.	☐	☐	☐	☐	☐
Durch das Tool stehen mir zusätzliche Informationen zur Verfügung.	☐	☐	☐	☐	☐
Das Tool war schwierig zu bedienen.	☐	☐	☐	☐	☐
Ich konnte die Verwendung des Tools gut in meinen klinischen Alltag integrieren.	☐	☐	☐	☐	☐
Durch das Tool ist meine Arbeitsbelastung gestiegen.	☐	☐	☐	☐	☐
Ich habe den Output des Tools in meine klinischen Entscheidungen miteinbezogen.	☐	☐	☐	☐	☐
Ich glaube, dass mithilfe des Tools ein Delir frühzeitig erkannt werden kann.	☐	☐	☐	☐	☐
Die angezeigten Informationen zu den einzelnen Patienten waren verständlich.	☐	☐	☐	☐	☐
In der täglichen Routine werden viele Delir-Fälle erst spät erkannt.	☐	☐	☐	☐	☐
Ich habe das Tool regelmäßig verwendet.	☐	☐	☐	☐	☐
Wie oft haben Sie das Delir-Prognose Tool verwendet (Ampel-Hinweis und/oder Detailanzeige)?	ca. _____ Mal pro Monat				
	Sehr selten	**Selten**	**Manch-mal**	**Häufig**	**Sehr häufig**
Wie häufig stimmte das berechnete Delir-Risiko mit Ihrer eigenen Einschätzung überein?	☐	☐	☐	☐	☐
Wie oft schätzten Sie selbst das Risiko höher als vom Tool angezeigt?	☐	☐	☐	☐	☐

Sonstige Kommentare/Anmerkungen:

Vielen Dank für Ihre Teilnahme!

Figure A.5: Questionnaire used to assess technology acceptance of the tool in German (original version).

EVALUATION – PILOT PROJECT DELIRIUM PREDICTION

To increase the quality of the delirium prediction application a many-sided evaluation is part of the pilot project. With this questionnaire, your experience with the delirium prediction application will be assessed. All your answers will be treated confidentially and cannot be traced back to you.

In case we need to combine different parts of the evaluation, a personal identification code will be generated.

1 2	3 4	5 6	
First and last letter of your father's given name (e.g. **K**ar**l** → KL)	First and last letter of your maternal grandmother's given name (e.g. **J**uli**a** → JA)	Last two numbers of your mother's year of birth (e.g. 19**48** → 48)	*In case you cannot provide an answer, please fill in the numbers 99.*

Job description: ☐ Physician in training ☐ Physician ☐ Nursing staff
Gender: ☐ male ☐ female
Age:

Please select for each statement the category which fits best for you.

	Strongly disagree	**Disagree**	**Neither agree, nor disagree**	**Agree**	**Strongly agree**
The purpose of the application was clear and understandable.	☐	☐	☐	☐	☐
The application is useful for my work.	☐	☐	☐	☐	☐
At time of implementation I was sufficiently prepared to use the application.	☐	☐	☐	☐	☐
I believe that the application is a useful support to prevent delirium.	☐	☐	☐	☐	☐
The application provides me with additional information.	☐	☐	☐	☐	☐
The application was difficult to use.*	☐	☐	☐	☐	☐
I successfully integrated the application into my clinical routine.	☐	☐	☐	☐	☐
The application has increased my workload.*	☐	☐	☐	☐	☐
I considered the output of the application in my clinical decisions.	☐	☐	☐	☐	☐
I believe that the application can be used to detect delirium at an early stage.	☐	☐	☐	☐	☐
The information displayed on the individual patients was understandable.	☐	☐	☐	☐	☐
In clinical routine, many cases of delirium are being detected only late.	☐	☐	☐	☐	☐
I have been using the application regularly.	☐	☐	☐	☐	☐

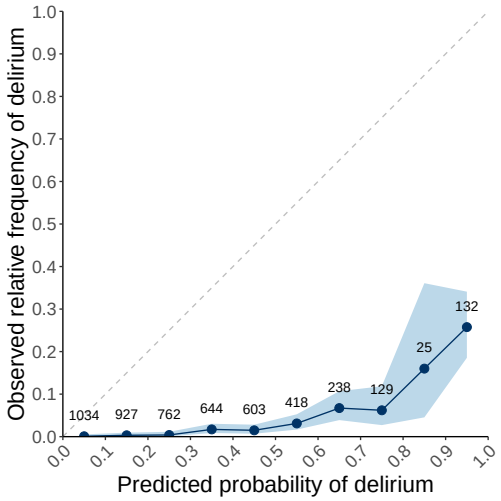
How often did you use the delirium prediction application? (Symbol in HIS and/or Web App?) around __ times per month

	Very Rarely	**Rarely**	**Some-times**	**Frequently**	**Very frequently**
How often did the presented delirium risk match your own risk estimation?	☐	☐	☐	☐	☐
How often did you estimate the risk to be higher than predicted by the application?	☐	☐	☐	☐	☐

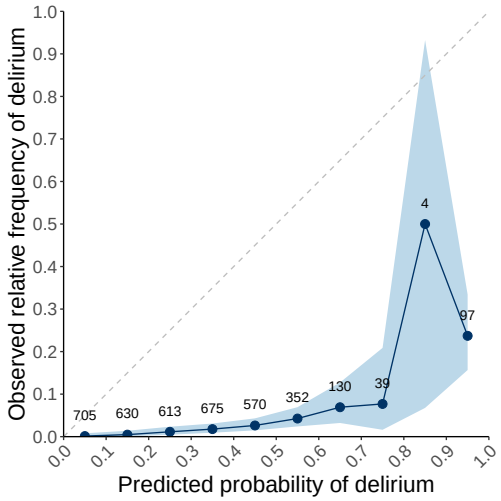
Comments:

Thank you for your participation!

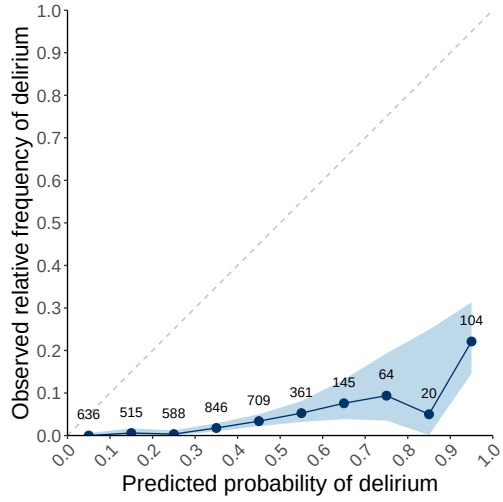
Figure A.6: Questionnaire used to assess technology acceptance in English (forward translation).



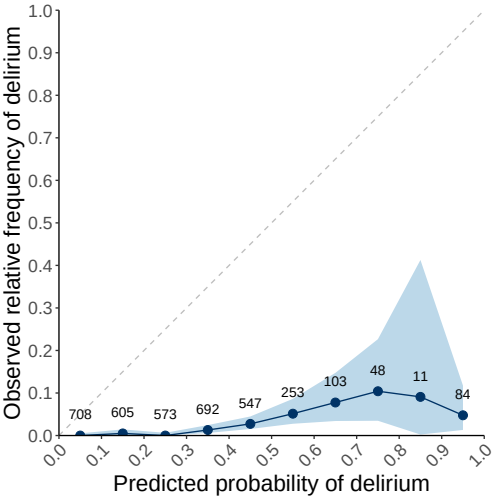
(a) Hospital 1 (Jan.-June)



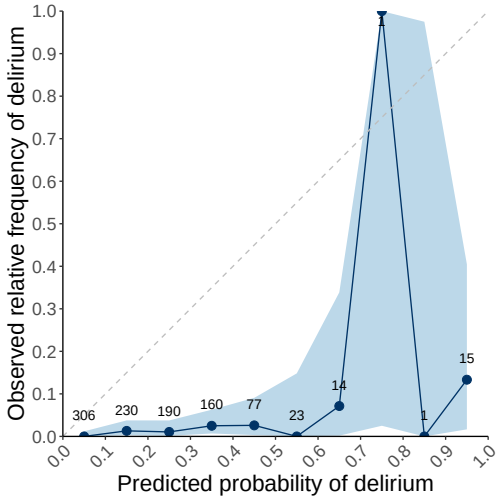
(b) Hospital 1 (Aug.-Dec.)



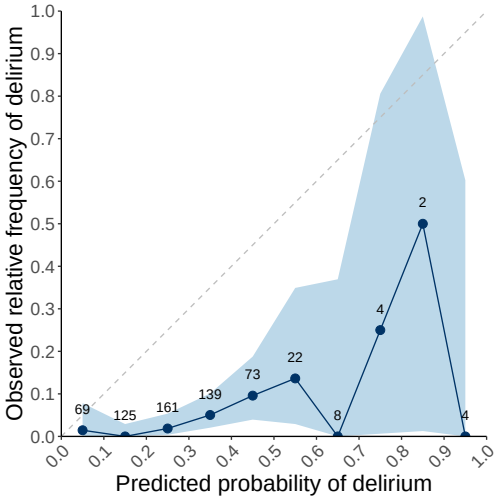
(c) Hospital 2



(d) Hospital 3



(e) Hospital 4



(f) Hospital 5

Figure A.7: Calibration plots for five hospitals during the long-term evaluation. Results for Hospital 1 were analysed separately (a) from January until June 2019, and (b) from August until December 2019.

B

TABLES

Table B.1: Search strategy to identify relevant literature describing the discriminative performance of delirium prediction models.

Research question	What is the discriminative performance of delirium prediction algorithms for inpatients (using electronic health records)?
Start date	01.01.2010
End date	31.12.2019
Literature sources	Pubmed
	EMBASE
Inclusion criteria	Discriminative performance reported in AUROC and/or sensitivity and specificity
	Delirium prediction for inpatients
Exclusion criteria	Full text not available
	Articles written in any other language than English or German
	Articles not peer-reviewed
	Studies without separate test data set or validation data set
	Studies including paediatric patients
Search terms	prediction OR predictive modelling OR machine learning OR random forest OR predictive model OR risk predict*
	AND
	delirium OR acute confusional state
	AND
	electronic health record OR EHR OR electronic medical record OR EMR OR clinical record OR information system
	AND
	admission OR hospitaliz* OR hospitalis* OR inpatient* OR in-patient*
	NOT children NOT pediatric NOT (ICU OR intensive care)

Table B.2: Description of feature groups of the two random forest models F05 and F10.4, including number of features, non-ranked examples as well as scale types (adapted from Jauk et al. (2020)).

Group	Features (n) F05	Features (n) F10.4	Examples (non-ranked)	Scale
Demographic	28	28	Age (in years)	int.
			Sex	nom.[1]
			Language German	nom.[1]
			Existing risk factors	nom.[1]
			Denomination	nom.[2]
			Academic degree	nom.[2]
			Marital status	nom.[2]
			ZIP Code	nom.[2]
			Private insurance	nom.[1]
Administrative	9	9	Number of transfers	int.
			Number of previous inpatient and outpatient stays	int.
			Number of previous inpatient stays	int.
			Admission from external hospital	nom.[1]
			Admission through emergency department	nom.[1]
			Longest inpatient stay (days)	int.
			Time since previous hospital stay (days)	int.
			Procedure-oriented hospital financing	int.
			Clinical department at hospitalisation	nom.
Diagnoses	309	174	Number of diagnoses	int.
			Charlson Comorbidity Index	ord.
			ICD 10 codes (three-digits code) e.g. E11 (Type 2 Diabetes mellitus), F00 (Dementia)	nom.[1]
			Number of coded ICD 10 codes within groups e.g. ICDgrp_E10_E14 (Diabetes mellitus), ICDgrp_I10_I15 (Hypertensive diseases)	int.
Laboratory	53	50	Abnormal laboratory values e.g. Bilirubin, Albumin, Gamma glutamyl transferase, Potassium, Mean corpuscular haemoglobin, Alanine aminotransferase	ord.
Procedures	94	75	Number of procedures (total)	int.
			Number of medical procedures	int.
			Number of nursing procedures	int.
			Number of procedures according to hospital specific catalogue e.g. Hip surgeries, Dialysis procedure, Computer tomography, Magnetic resonance diagnostic	int.
Nursing	91	89	General condition e.g. Respiratory problems, use of sedatives, smoking, falls, hearing aids, glasses, incontinence	nom.[1]
			Independency levels e.g. bathing, nutrition, hydration, mobility	ord.
			Body Mass Index	int.
			Orientation level at admission	ord.

Note: [1]dichotomous variable; [2]several dummy variables were computed for this feature.

Table B.3: Frequency of EHR records for delirium in five KAGes hospitals. Patients with delirium were identified by ICD-10 codes and text mining of discharge summaries.

Hospital	N	ICD-10 code		Discharge summaries		Discharge summaries only	
		n	%	*n*	%	*n*	%
Hospital 1	221	83	37.6	213	96.4	138	62.4
Hospital 2	105	39	37.1	101	96.2	66	62.9
Hospital 3	60	16	26.7	58	95.0	44	73.3
Hospital 4	15	8	53.3	14	93.3	7	46.7
Hospital 5	23	7	30.4	23	100.0	16	69.6
Total	424	153	36.1	409	96.5	271	63.9

Table B.4: List of stop words excluded during the cleansing of discharge summaries of delirium patients. Stop words were manually defined in an iterative process.

°c, abklärung, absetzen, abteilung, anamnese, aorta, art, ätiologie, aufenthalt*, aufgenommen, aufgrund, aufnahme, äußeren, bauchdecke, bds, befund*, beginn, behandelt, behandlung, beidseitig, bekannter, bereich, bereits, besserung, besteht, bzw, degenerative, deutlich, diagnose*, durchgeführt, dzt, eigenanamnese, eingeleitet, einsehbar, empfehlen, empfohlene, entlassen, entlassung, entsprechend, entwickelte, entwicklung, erfolgt*, erster, etabliert, fehlender, folge, frau, fußpulse, gehirnschädels, genese, geringe, graz, groß, gut, harnblase, hauptdiagnose, herr, herrn, iel, jedoch, kam, klinischer, kommt, konnte, konsil, linie, links, lkh, mehr, mittels, möglich, neurologischer, nieren, non, norm, novalgin, oben, pat, patient*, projektion, prompt, rachen, rahmen, raumforderungszeichen, rechts, regelrecht, relevante, resistenzen, rund, schädel, schließlich, schmal, schmerzen, sei, seit, sodass, sowie, spätem, standort, stat, station*, status, stuhl, süd, symptomatik, tagen, täglich, temp, therapie*, transferiert, tropfen, typ, übernahme, univdoz, unklarer, unser*, verabreicht, veränderungen, verdacht, verlauf, verschlechterung, version, verzichtet, vorerkrankungen, vorlage, vorstellung, vorübergehend, wäre, wegen, weitere*, weststeiermark, wiedervorstellung, wirbelsäule, wurde, zeigt*, zuletzt, zunehmend*, zusammenfassung, zusätzlich, zustand*