Leveraging AI-based prediction in perioperative and critical care: from model development to clinical implementation

Meijden, S.L. van der

Chapter 9

General Discussion

This dissertation's overall aim was to investigate the potential of AI to be of clinical value, specifically in perioperative and critical care, along the phases of the AI development and evaluation trajectory. Nonetheless, the findings of this research can be adopted outside the perioperative and critical care domain, and its broader implications will be discussed. I reflect on 1) the human and data-related challenges encountered in this dissertation, 2) how to establish valid and fair use of AI-based prediction models, and 3) the effectiveness and performance in predicting patient outcomes, specifically when compared to physicians. Lastly, a future outlook and research recommendations are provided.

9.1. Human and data-related challenges for integrating AI into clinical practice

Introducing AI tools into clinical practice comes with a broad range of challenges related to human factors (**chapter 3**), model explainability [1], data and label quality (**chapter 4**), model generalizability (**chapter 6**), ethics and fairness (**chapter 8**), the implementation process [2], legislation and regulation [3], and information technology infrastructures [4]. In this section, I discuss the contributions of this dissertation to addressing human factors and data-related challenges, specifically on how these challenges apply to PERISCOPE. PERISCOPE was the main use case throughout this dissertation, predicting the risk of postoperative infection within 7- and 30-days of discharge that will be integrated into the electronic health record (EHR).

9.1.1. Human factors: Physicians' perspectives on using AI in clinical practice

The majority of current available AI tools aim to provide decision support to a certain 'intended user' and will not work autonomously over the coming years [5]. These intended users, in this dissertation mostly healthcare professionals, will therefore interact with the AI tool to be of support in their decision-making. Several human factors should be considered to achieve effective interaction between healthcare professionals and AI tools. A review of human factors influencing the interaction of healthcare professionals and AI identified five relevant domains: 1) medical expertise, 2) technological expertise, 3) personality, 4) cognitive biases, and 5) trust [6]. To address these aspects of the intended user group in the pre-implementation phase and to investigate how the AI tool should be best integrated into the clinical workflow, end-users should be closely involved throughout the process. In **chapter 3**, intensive care unit (ICU) physicians from two hospitals were questioned on their perspectives on the future use of an AI tool predicting ICU readmission risk. A promising 86% (n=55) of respondents believed that AI could support them in their work as a physician and that they were willing to incorporate AI in their clinical practice.

However, physicians at the site where the model was developed were more familiar with AI and more in favor of the ICU readmission prediction tool compared to the physicians from the non-development site. This finding may imply that the adoption of AI tools in non-development centers and middle-to-lower-income countries may be more challenging due to a familiarity- and knowledge gap. To ensure that AI tools are accepted and used safely by their intended users, there is a need to educate clinicians and ensure that the right tools are developed for the right

questions [7]. To effectively advance clinicians' knowledge concerning the opportunities and pitfalls of using AI in healthcare, I see two crucial steps: 1) including foundational AI courses in medical curricula [8] and 2) publishing comprehensive overview papers for each clinical domain (see for an example **chapter 2**).

A term often used when discussing barriers and facilitators for adapting AI tools into clinical practice is 'trustworthiness' [9, 10]. Just as people need to trust the cars they drive or the apps they use for money transfers, clinicians must have confidence in AI tools before allowing them to influence their decision-making. The multi-faceted issue of trustworthiness involves ethical, social, and technical considerations. Terms that are repeatedly linked to trustworthiness are AI interpretability and explainability. While often used interchangeably, the distinction has been made that AI interpretability refers to human understanding of the output and working mechanisms of the AI tool, whereas AI explainability refers to techniques used to explain predictions made by complex, 'black-box' AI tools [11]. In **chapter 3,** respondents indicated that they want to be able to see on which patient factors a prediction is based, indicating the need for AI explainability. Many methods have been developed to provide AI model 'explainations' i.e., show the most important predictors per patient [12]. However, using explainable AI may introduce the false idea that there is causality instead of a correlation between predictors and the outcome [13]. If not instructed properly, this may lead to the acting of clinicians to predictors that are not causally related to the outcome (e.g., *a patient has a high predicted risk of infection ⇢ model explanation shows that the high blood pressure of a patient is correlated to the increased risk of infection ⇢ clinician falsely interprets that the blood pressure should be lowered ⇢ prescribed blood pressure lowering medication*). Explainable AI should therefore be used with caution. Some even state that *"we should stop explaining AI models for high stakes decisions and use interpretable models instead"* [14]. I argue that explainable AI may be of value in determining the clinical basis supporting the prediction and discovering potential errors in input data, such as done in **chapter 6**. However, this requires the training of clinicians and providing 'warnings' in user interfaces that explain the concepts behind it.

9.1.2. Data: Leveraging AI by using routinely collected electronic healthcare data

On the technical side are the challenges associated with using EHR data for training and validating AI-based prediction models, as well as for making predictions after the AI tool is deployed. Unlike evidence-based medicine, which relies on prospective randomized controlled clinical trials to conclude, AI predominantly uses retrospective observational patient data to attain sufficient sample sizes for model development [15, 16]. By learning from thousands to millions of patient records, AI-based prediction models can utilize this vast amount of data to predict outcomes across a broad spectrum of patients. However, poor data quality is more the norm than the exception, often resulting in biased predictions. This phenomenon is commonly referred to as "garbage in, garbage out" [17, 18]. I want to emphasize the importance of evaluating data quality before starting modeling endeavors, for example using the recently published METRIC framework [19].

Due to the heterogeneity in data collection and storage methods between hospitals, scaling AI tools across hospitals requires manual labor and limits transferability. For example, different coding systems for medication use may be used. The increasing adoption of common data models such as Fast Healthcare Interoperability Resources (FHIR) and Observational Medical Outcomes Partnership (OMOP) is enhancing data transferability [20, 21]. However, for the scaling of PERISCOPE across FHIR hospitals we encountered that manual data mapping was still required. With the rise of generative AI and large language models, this process could be supported, for example by selecting the appropriate diagnostic codes for postoperative infections [22]. Furthermore, adhering to the FAIR (Findable, Accessible, Interoperable, Reusable) principles would enable a more standardized, interoperable data environment enabling smoother and more efficient scalability of AI tools like PERISCOPE across healthcare systems.

Another significant challenge in using observational data for AI is the lack of reliable labels to train and validate models on. For the PERISCOPE AI tool, we needed to identify (i.e., label) patients with postoperative infections based on retrospective, observational EHR data. A reliable labeling method aiming to be close to the 'ground truth' is essential to 1) let the model learn the correct relationships between input data and the outcome of interest, 2) validate model performance by comparing the prediction to the label, and 3) monitor real-life model performance once deployed in the hospital (Figure 1).

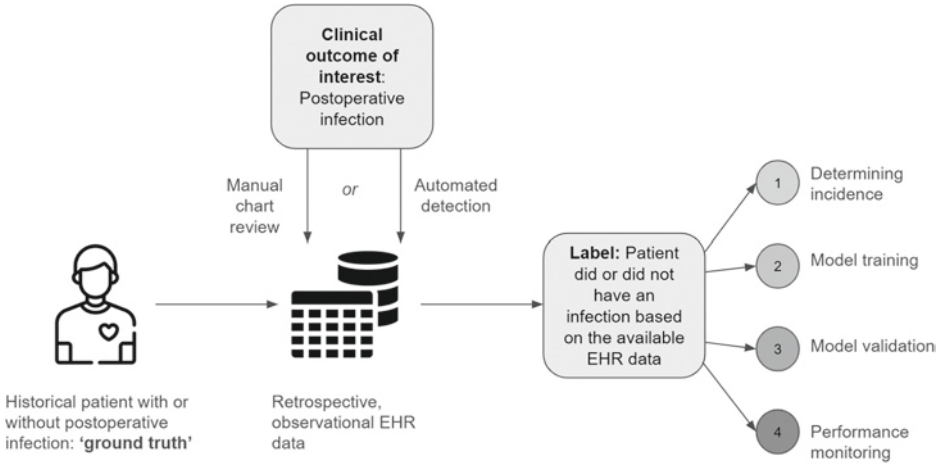

Figure 1. Illustration of how labels used in predictive modeling can be derived automatically or by manual chart review from retrospective, observational electronic health record (EHR) data. The labels should represent the 'ground truth' as accurately as possible to prevent errors in determining baseline incidences, model training, model validation, and performance monitoring.

Complication registries are frequently used for labeling, surveillance, and monitoring quality improvement efforts [23]. However, the high underreporting rate of complications highlights the need for alternative methods [24, 25]. Performing manual labeling through EHR chart review is standard practice for many AI tools. The limitations of manual labeling are discussed in **chapter 4**, where we reviewed automated labeling methods to identify patients with postoperative

infections. We learned that there are no unified definitions for postoperative infections and that only a few prediction modeling studies used alternative methods to manual chart review for labeling. The latter contributes to the lack of scalability of AI tools as it is infeasible to perform manual chart reviews for thousands to millions of patients per hospital.

For PERISCOPE, we collaborated with surgical, infectious disease, and intensive care specialists to develop a clinically relevant definition of infections that allowed automated labeling using EHR data. This comprehensive definition, which included complication registrations, surgical interventions, and pharmacological treatments, was validated against human chart reviews in **chapter 5**. It is important to recognize that achieving 100% accurate labeling from observational data is nearly impossible. For example, even with outcomes like mortality, patients who return to their home country may pass away there, and these deaths may not be recorded in the EHR of the country they initially visited. The absence of reliable 'ground truths' is a pervasive challenge in medical research, not just in AI model development, often due to inter- and intra-rater variability in diagnoses. To ensure reliable labeling, I propose the following steps: 1) involve clinical and data experts, along with established guidelines, to define the outcome of interest and corresponding criteria, 2) assess multicenter EHR data for completeness and quality regarding the defined criteria, 3) iteratively refine the definition with clinical and data experts, and 4) validate the definition against manual chart reviews, acknowledging that manual review is also susceptible to errors.

9.2. Valid and fair use of AI-based prediction models across various clinical settings

When scaling AI tools to different clinical settings and hospitals, it is crucial to ensure that they are locally valid. This means that the tool must make accurate predictions for the specific hospital population. Additionally, it is important to ensure that the tools are fair, avoiding the exacerbation of health disparities among marginalized groups. In this dissertation, the steps towards locally valid and fair AI tools were studied. First, the PERISCOPE AI tool was developed and internally validated as a proof-of-concept at the Leiden University Medical Center (**chapter 5**). Second, the PERISCOPE AI tool was externally validated and locally updated in two additional hospitals (**chapter 6**). With the growing use of AI tools in decision-making, model fairness has become a focal point in debates on healthcare disparities, particularly regarding how biased data and biased predictions may impact marginalized patient groups. However, this field is complicated with numerous definitions and evaluation methods to assess fairness (**chapter 8**). The steps towards valid and fair clinical AI prediction models are discussed in this section. Additionally, I address the certification and regulatory issues involved with the use of AI tools in clinical practice.

9.2.1. How do we define the valid use of AI tools?

In a perfect world, although unrealistic, the following equation would hold;

$$\Sigma_{i=0}^{N} \quad |\hat{y}_i - y_i| \quad = 0$$

where $\hat{y}_i$ is the predicted class [0,1] and y_i the true outcome (label) of a patient *i*. As no model can predict perfectly, the predictive performance in terms of discriminative performance, calibration, and clinical utility should be assessed during model validation (see **chapter 1**, Introduction, Figure 3). The term 'model validation' is used throughout the literature on prediction model development. However, the term validation is used inconsistently, sometimes referring to tuning the model, tuning, and testing the model, or only testing the model [26]. The distinction is often made between internal and external validation [27], or further categorized as internal, internal-external if several hospital datasets are used during development, and external validation [28, 29]. During internal validation, model performance is assessed on the development dataset to determine the risk of overfitting and reproducibility of results [30]. During external validation, model performance is evaluated using unseen data, either from different time periods or geographical locations, to assess its generalizability and transferability to other clinical settings. However, achieving a truly 'validated prediction model' may not be possible [30]. Temporal and geographical variations in patient populations, novel therapeutic strategies, and data measurement procedures mean that even after external validation, there remains uncertainty about how a model will perform in a new setting [31].

While the goal is still often to achieve 'global validity', i.e., a model that can generalize across various settings without the need for local adjustments or further validation, findings from **chapter 6** highlight the importance of focusing on 'local validity'. This approach accounts for 'domain shift,' which includes differences in case mix, disease incidence, treatment pathways, and data collection methods.

9.2.2. Validation and updating steps

As stated in the previous paragraph, the truly 'validated model' does not exist. In this dissertation, we investigated the validation and updating steps to allow for assessing local validity before implementing an AI-based prediction model in a clinical setting (**chapter 6**). Implementing AI-based prediction models without local validations can result in a dramatic drop in performance. This was sadly illustrated by a widely implemented sepsis prediction model, leading to a high rate of false negatives and false positives [32]. While local validating and updating of AI-based prediction models are resource intensive, I see this as a vital step for currently available AI tools like PERISCOPE. Acquiring a sufficient sample size and high-quality data for validation and model updates can be challenging [15]. However, these challenges must be addressed when planning to implement a model, as high data quality is also essential for successful deployment. Failing to assess data quality and model performance before implementation increases the risk of performance degradation.

We distinguished the validation and updating steps for AI-based prediction model development: phases 1 (AI model development) and 2 (assessment of AI performance and reliability) before going live in clinical practice (Table 1). Once deployed, it is vital to monitor model performance to account for changes in input data (i.e., data drift) and assess the need for model retraining or recalibration [33]. To allow for global models that can be adapted to site-specific needs, other options like membership models and federated learning could be explored [34, 35].

Table 1. Model validation and updating steps in phase 1 and 2 of the model development and evaluation trajectory

Phase	Site	Validation and updating steps	Explanation	Goal
AI model development and validation	Development site	1. Internal validation	Training and validating the model on different subsets of the development dataset (cross-validation) or using resampling techniques (bootstrapping)	Evaluating model robustness and risk of overfitting
		2. Temporal external validation	Splitting the train and validation dataset based on time, where the most recent dataset is used for validation	Evaluating generalizability over time
Assessment of AI performance and reliability	Implementation site	3. Geographical external validation	Applying the model directly to all available data from the implementation site	Evaluating the generalizability of the model to a new clinical setting and determining the need for model updating
		4. Model updating	Based on the geographical external validation results, either retraining and/or recalibrating the model. Use a temporal split for creating an updating and validation dataset.	Optimize performance and adapt the model to a local setting
Assessment of AI performance and reliability	Implementation site	5. Local validation	Evaluate the final, updated model as done in step 1.	Assess final model performance
		6. Prospective 'underwater' validation	Evaluate the final model 'underwater' (ie, not shown to the end-user) on live data before implementation in the clinical setting	Assess model robustness on live incoming data that may differ in completeness and quality compared to retrospective data

Figure 2. Pubmed results from 1983-2024 for "*fairness*" AND "*prediction*"[1].

1 Available on: pubmed.ncbi.nlm.nih.gov/?term=prediction+AND+fairness&filter=years.1983-2024

9.2.3. Fairness

Besides the validity of AI-based prediction models, the concept of 'fairness' has become a debate for using these models in clinical practice (Figure 2). It seems that with the rise of AI-based prediction models, fairness gained more attention compared to 'classical' predictive models that have already been developed and in use for the last decades. In **chapter 8**, we aimed to structure the transition of AI fairness from theory to practice, as the field is complicated with numerous definitions and evaluation metrics. Twenty-seven different fairness evaluation metrics were identified, many of which are incompatible with one another and may not be statistically or epidemiologically appropriate for the medical domain [36]. I see the rising attention to fairness over the past few years three-folded: 1) AI-based or informed decision-making is becoming integrated into our society, which may cause more harm than good to certain minority groups, 2) there is an increased awareness of (racial) biases within and outside the medical domain (e.g., the "toeslagenaffaire" in the Netherlands [37]) that may be exacerbated when building predictive models based on a biased dataset, and 3) for 'black-box' AI-tools that include protected attributes such as Race or socioeconomic status, it is for end-users impossible to determine on what, potentially biased factors a prediction is based.

The fundamental problem in ensuring fair AI-based decision-making, is the biases present in the data used for developing the algorithm [38]. For example, a sepsis prediction algorithm is developed for the ICU. Due to the underrepresentation of patients from certain demographic groups, the AI tool might perform well in the majority group (e.g., White men) and perform worse in other groups (e.g., Black women). This could lead to delayed diagnoses in the minority group, and therefore increase health disparities. Another example is the use of proxy labels to predict patient outcomes. 'Postoperative infection diagnosis' is a proxy for postoperative infection incidence. Underdiagnosing in certain groups may result in algorithmic biases, and consequently unfairness.

In medical applications, patients often benefit most from an accurate prediction and not a higher or lower prediction which is often true for non-medical applications. This so-called 'non-polar' decision-making in the medical domain results in most fairness metrics being irrelevant [39]. In some applications, it may however be worth performing bias mitigation measures to reduce health disparities. For example, to allocate scarce resources to marginalized patient groups. Furthermore, I recommend using the framework from **chapter 8**, addressing discriminative performance, calibration, and clinical utility, to evaluate whether the model underperforms in certain groups and to evaluate the effect of bias mitigation measures. AI models may perform worse in minority groups, which can only be identified when addressing model performance in these groups [40, 41]. However, before evaluating model fairness, it is fundamental to assess and address biases in the training data [42]. The complexity of the effort to address unfairness is that patients can belong to multiple patient groups and can be of, for example, mixed race. How to incorporate this in the fairness debate should be further investigated.

9.2.4. Certification of AI models: enhanced safety or blocked innovation?

To guide and guard the safe use of AI tools in clinical practice and beyond, developers need to comply with the applicable regulations for their region. As PERISCOPE was developed to be used as a 'software as a medical device', specific certifications needed to be acquired [43]. The research performed in this dissertation (**chapter 4, 5 and 6**) was used as clinical evidence to obtain CE certification under the Medical Device Regulation (MDR) [44]. The CE certificate was a requirement to sell PERISCOPE as a medical product to European hospitals and therefore essential to allow use outside the research setting [45]. Over five years of research, development, and quality assurance work resulted in the CE certification of PERISCOPE. The work of this dissertation mostly contributed to the 'clinical evaluation' part of PERISCOPE's technical documentation, where we established the 'benefit-risk ratio' (i.e., the potential benefits of using PERISCOPE outweigh the risks) and whether it was suitable for its intended purpose. Furthermore, to ensure the digital safety of PERISCOPE, both ISO13485 certification for quality management systems and ISO27001 for cybersecurity certification were obtained. Additional regulations must also be considered, such as the GDPR in Europe and its less stringent counterpart, HIPAA, in the United States [46, 47]. The recently enacted European AI Act introduces new regulations for 'high-risk' devices, which include all medical AI tools [48]. For PERISCOPE to enter the US market, it will also need to secure approval from the Food and Drug Administration (FDA).

This abundance of legislation developers of AI tools have to comply with, is causing a major hurdle in getting AI tools to market as medical devices [49]. There seems to be a trade-off between innovation speed and ensuring safety, where start-ups need several rounds of funding and a high level of expertise to even kick off deployment. Once the required certifications are achieved, post-market surveillance should be performed which means that the performance and potential adverse events must be monitored. I estimate that 20-40% of start-up resources (financial, employees, time) are spent on developing and maintaining a quality management system and acquiring the required certifications. While these regulations are essential to ensure the safety of new and existing AI tools intended to be used in clinical settings, the current system is blocking innovation. I recommend investing in one comprehensive and comprehensible guideline where all the requirements and steps are detailed and outlined to be followed by AI developers in healthcare. Dedicated notified bodies, the instances that audit and determine whether a company complies with the legislation, should be assigned to allow and assist in the certification process.

9.3. AI effectiveness and performance

For AI tools to be of value in perioperative and critical care, they should be effective (i.e., high-performing) in predicting patient outcomes. In this section, I reflect on PERISCOPE's ability to predict postoperative infections (**chapter 6**). Furthermore, the comparison between PERISCOPE and physicians made in **chapter 7** is discussed. This dissertation does not examine the direct impact of AI tools on decision-making and patient outcomes. Several implementation challenges were encountered that delayed or hindered the integration of AI tools like PERISCOPE into

clinical practice. Therefore, I address the implementation challenges and the clinical trials required to evaluate the effectiveness of AI tools within real-world healthcare settings to go from development to clinical usefulness.

9.3.1. Performance of models in perioperative and critical care: global versus local models

Following the steps in Table 1, PERISCOPE achieved, after updating, a discriminative performance (area under the receiver operating characteristic curve (AUROC)) › 0.8, a well-fitted calibration curve, and a net benefit higher than default strategies in three distinct hospitals (**chapter 6**). Based on these results, we deemed PERISCOPE's performance sufficient to allow for clinical implementation. However, there are no hard cut-offs to determine if an AI tool is suitable for clinical use, where there is heterogeneity in when the AUROC is perceived as 'good' [50]. To evaluate PERISCOPE in the CE certification process, we established that the average predictive performance for postoperative infections, across both statistical and AI models, was an AUROC of 0.7. We established this as our performance benchmark as PERISCOPE's performance should exceed the current standard. PERISCOPE will therefore not be implemented for specialties not meeting this benchmark based on local validation results.

Systematic reviews on AI-based prediction models in surgical and critical care found that 80.6% (n=29) and 85.2% (n=421) of studies, respectively, relied solely on internal validation with small datasets [51, 52]. It is time to move beyond publishing just the internal validation results of AI-based prediction models intended for clinical use, as these results offer limited scientific and practical value. I recommend conducting at least one external validation with an adequate sample size, with or without model updating, before publishing the results [53]. What an adequate sample size is depends on several factors, including the anticipated outcome event rate [54]. Furthermore, reporting standards like the TRIPOD-AI and DECIDE-AI guidelines should be followed to ensure complete reporting of performance outcomes (discriminative performance, calibration, and clinical utility) and other important study characteristics [55, 56].

PERISCOPE's updating approach for creating local models is also seen for other applications, such as intensive care unit readmission prediction [57]. Updating models to create local versions is slowly becoming more standard practice [58], in contrast to previously developed, often classical statistical predictive models. When developing predictive models, we aim to capture the underlying 'disease model', which should be transferable between sites when applying the model to a similar population [59]. For example, suppose a model is trained and validated to accurately predict infection risk at site A, which includes a large and diverse patient population and high-quality predictors. In that case, it should ideally capture transferable relationships between the predictors and the outcome of postoperative infection. We would then expect these relationships to hold when the model is applied at site B. However, we often observe a drop in performance when transferring a prediction model to a new site. This 'failure to transport' predictive models between sites can be accounted for by the 'process model' representing the site-specific changes in data collection methods, patients, and treatment pathways [59].

Others have distinguished the types of uncertainty causing a drop in performance as aleatoric and epistemic uncertainty [60]. Aleatoric uncertainty refers to the stochastic or random nature of factors influencing whether an outcome occurs or not. Aleatoric uncertainty will always remain regardless of how much information the model has about the underlying phenomenon. For example, in two patients with the same characteristics, one will get an infection and the other does not. The other type of uncertainty is epistemic uncertainty, which is reducible by gathering more data or improving the model's architecture. When there is high epistemic uncertainty, for example, due to differences in local data acquisition methods, the generalizability of the model is often affected. Local models are tailored to the specific patient population, potentially reducing epistemic uncertainty in those areas by focusing on more relevant data. However, local models can still suffer from epistemic uncertainty if the sample size of local datasets is sparse.

9.3.2. Doctor versus algorithm

To assess whether end-users may benefit from AI-based prediction tools, comparisons between human prediction and predictive algorithms have been made [61]. For PERISCOPE, we questioned 51 surgeons (in training) from various specialties to predict the risk of infection within 30 days of surgery and compared them to the predictions of PERISCOPE (**chapter 7**). 544 predictions were collected, for which PERISCOPE performed on par with the clinicians in terms of AUROC. However, when surgeons expressed uncertainty in their predictions, PERISCOPE had higher performance. Notably, while most participants were experienced and skilled surgeons, PERISCOPE also surpassed the performance of the younger surgeons-in-training involved in the study. However, the sample sizes of these subgroups were small, where further research is needed to investigate the potential of PERISCOPE to be of benefit in decision-making. Discussing the results of this study with the participating surgeons, they indicated that they made a thorough assessment before making their estimate and that it would be helpful for them to have a readily available prediction to support their decision-making. The next crucial step is to evaluate the 'doctor + the algorithm' to determine the influence on decision-making and ultimately on patient outcomes.

9.3.3. Implementation challenges for integrating AI into clinical practice

Compared to other clinical domains like radiology and pathology, the uptake of AI tools in perioperative and critical care remains slow [8, 62]. The literature has widely discussed the barriers and facilitators for implementing AI tools in clinical practice [2, 8, 63, 64]. In **chapter 2**, we identified several barriers to the safe implementation of AI, including variations in study quality that can lead to potential biases and reduced generalizability, as well as technological, regulatory, explainability, and data-related challenges. This is not an extensive list as ethical, financial, social, and legal factors also play an important role [64]. The acknowledgement of these challenges is important to go from research projects to real-life deployment. However, I want to emphasize that over the past few years, there has been a noticeable shift in the attitudes of hospitals, EHR providers, clinicians, and governments toward AI implementation. Additionally, these stakeholders have increasingly allocated more resources to support the

integration of AI in healthcare. Where in 2021 mostly academic hospitals were investigating the potential of AI in clinical care, an increasing number of non-academic hospitals are currently enabling AI implementations by dedicating resources [65, 66]. Multidisciplinary teams within hospitals are essential to allow successful implementation. I recommend involving a project leader, the IT department (including EHR and data (science) experts), the financial department (to determine the business case), the legal department, and the medical staff that will be the end users. Optionally, the ethical department, patient organizations, and medical technology department could be of value to enhance safe implementation.

To enhance the clinical adoption of AI tools, implementation research is of major importance [8]. Implementation research focuses on eight different outcomes: 1) acceptability, 2) appropriateness, 3) feasibility, 4) fidelity, 5) adoption, 6) penetration, 7) implementation cost, and 8) sustainability [2, 67]. While implementation research was not the focus of this dissertation, the next step for AI tools like PERISCOPE is to study clinical impact as well as implementation outcomes**.** I distinguish challenges for effective implementation of AI tools in perioperative and critical care categorized into clinical and human factors, technical, and regulatory (Table 2). The regulatory and financial challenges are mostly focused on the development of AI tools by the industry (e.g., start-ups), as different regulations apply to in-house development.

Table 2: Implementation challenges per phase of the AI development and evaluation trajectory.

Phase	Clinical and human factors challenges	Technical challenges	Regulatory & financial challenges
Preparations prior to model development	• Determining the clinical outcome of interest that can be identified in EHR data for labeling and validating it	• Collect data of acceptable quality and sufficient sample size from EHR databases	• Acquire funding • Comply with relevant regulations (GDPR, MDR, FDA, AI act, etc.) before starting data extraction and exploration
AI model development	• Important risk factors for the outcome may not be available (of sufficient quality) in the data	• Ensuring that biases in data do not affect model fairness • Ensuring model explainability	• Data privacy (e.g., comply with, ISO27001)
Assessment of AI performance and reliability	• Determining when a model is of sufficient performance and is fair 'enough' to be safe to use in clinical practice	• Developing a user interface that is easy and self-explanatory to use • Transition from using the model on retrospective data to live data for prospective, underwater testing	• Determining potential cost-effectiveness • Acquire sufficient resources for certification processes (industry) and expertise within the hospital on relevant regulations

[continued on next page]

Table 2: *[continued]*

Phase	Clinical and human factors challenges	Technical challenges	Regulatory & financial challenges
Clinical testing of AI	• Designing appropriate studies to evaluate impact where randomized controlled trials are not always suitable • Evaluating implementation outcomes	• Implementing AI user interface in EHR	• Acquire funding to perform clinical trials
Implementing and governing AI	• Training of end-users • Effective integration in the workflow	• Scaling product to other hospital settings and datasets, including local data and model validations. • Live data may be of less quality than retrospective data • Monitoring of real-life performance and input data, and determining steps to take when performance drops	• Having a convincing business case • Arrange reimbursement from healthcare insurance companies • Complete certification process

9.3.4. From performance to impact

As described in the previous paragraph, implementation research is of major importance to go from statistically highly performing AI-based prediction models to clinical impact. Here, I further discuss the types of evidence that should be collected to assess the impact of AI tools on clinical outcomes (Figure 3). Conducting these studies will allow us to determine whether AI can truly make a difference in perioperative and critical care, areas where high complication rates remain a significant concern for patients, healthcare providers, and hospitals.

The highest level of clinical evidence is gathered through, preferably multiple, randomized controlled trials (RCTs) [68]. In the realm of evaluating AI tools for clinical purposes, 86 RCTs have been performed, of which only 6% (n=5) in perioperative care and 1% (n=1) in critical care [69]. Half of these studies focused on clinical outcome prediction, two on intraoperative hypotension prediction, and one on sepsis prediction [70-72]. While the results of these trials were positive, it is not possible to determine from these few studies whether AI-based prediction models have a high impact on patient care, as the outcomes evaluated mostly were focused on diagnostic yield or performance [69].

For RCTs evaluating the impact of AI-based prediction models, randomization can take place on a patient level, end-user level, or department level. The latter is often done through a so-called stepped-wedged cluster randomized trial design. Here it is randomized per department when the tool is implemented in a stepped approach (after a defined time step, one additional

department is implemented). Stepped-wedged trials are applied when individual randomization is not possible or not desirable [73], which is often true for AI tools that are integrated into the EHR. The different departments are exposed to the AI tool in a stepped approach to control for adoption bias and temporal changes in the background patient characteristics [74].

Cost-effectiveness or cost-utility analyses may help determine the potential for the AI tool to have a positive impact on patient outcomes and costs. Once AI tools are deployed beyond these research settings, it is important to monitor and evaluate the tool's effectiveness on patient outcomes and costs. Furthermore, a Health Technology Assessment (HTA) evaluating the clinical, economic, ethical, and social implications to determine its overall value and impact on healthcare delivery should be performed in this stage which can help get the AI tool costs reimbursed [75].

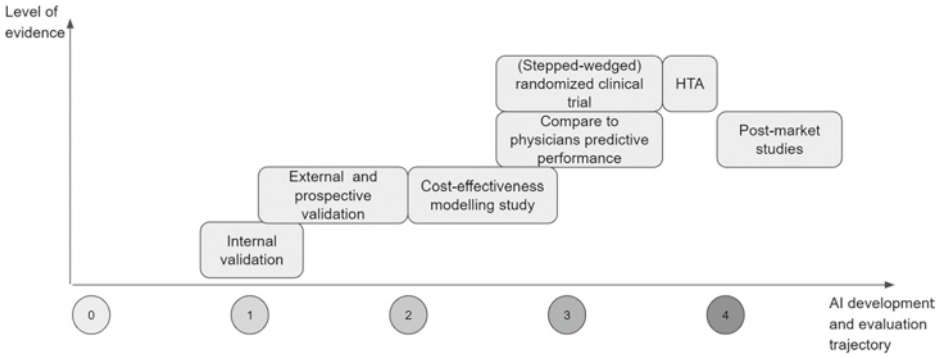

Figure 3. Studies conducted throughout the AI development and evaluation trajectory with corresponding levels of evidence. HTA = Health technology assessment.

For PERISCOPE, the next steps are to conduct a cost-effectiveness study, followed by real-life implementations. Stepped-wedged randomized clinical trials will be performed to evaluate the impact of PERISCOPE on various outcomes. Ideally, the quadruple aim should be met, including 1) reduction of costs, 2) improvement of population health, 3) improvement of patient experience, and 4) improvement of healthcare team well-being [76]. To achieve a substantial impact on these four domains we might need further product development in terms of for example providing tailored decision support on diagnostic and treatment decisions. The current version of PERISCOPE has, however, significant potential to support clinicians in their daily work by assisting with triage and reducing the time needed to assess a patient's overall infection status.

Lastly, I want to highlight that most of the currently available AI tools can only be applied in highly technologically advanced, high-income countries. This may increase global healthcare disparities, but there is also the opportunity for AI tools to assist in lower-resource settings in middle and low-income countries [77]. For example, AI tools have been developed to predict

deterioration in pediatric care units for use in Africa [78]. I see this as an important direction for AI-based prediction models to improve healthcare globally.

9.4. The future of AI in perioperative and critical care

In this section, I shortly elaborate on the prospect of AI in perioperative and critical care and beyond, before providing recommendations for future research.

9.4.1. Prospect

AI-based prediction models are gradually making their way into clinical settings. However, in the coming years, they are expected to serve primarily as decision-support tools or informational aids, rather than replacing clinicians in perioperative and critical care environments. In other domains, such as radiology, AI tools may already surpass human performance in diagnosing and treatment planning. However, the complex decision-making required in perioperative and critical care is expected to remain in the hands of humans for the foreseeable future. To achieve its potential, I outline the prospects for different aspects related to the clinical use of AI in Table 3.

Table 3: Prospect for effective use of AI tools in clinical settings

Domain	Prospect
Hospital	Each hospital will have a dedicated AI team focusing on 1) procurement and evaluation of existing AI tools, 2) policy and governance of using AI tools, 3) informing departments and healthcare professionals, and optionally 4) developing AI tools for in-house use in collaboration with clinicians.
Industry	Start-ups and other companies will collaborate closely with EHR vendors to ensure integration in the workflow and start certification processes as soon as possible.
Education	Training nurses and physicians as part of the standard curriculum for the safe use of AI tools in clinical practice.
Global Health	Expanding the use of AI tools to low resource settings, for example in diagnosis and early warning detection [78] to fight healthcare disparities [79].
Data	Complying with the FAIR principles (findable, accessible, interoperable, and reusable), including adhering to data standards (e.g., FHIR, OMOP [20]), will improve both data quality and transferability.
Certification	Notified bodies will be dedicated to the comprehensive certification process for AI tools to be employed as “software as a medical device”. Specific guidance is outlined to simplify the steps towards certification.
Insurance	Insurance companies will start reimbursing the use of AI tools when positive HTA results are reported
Ethics	Ethicists, patients, and end-users will be included throughout the model development process, where fairness is a central theme.

9.4.2. Recommendations for future research

The recommendations for future research categorized by the three research questions of this dissertation are summarized in Table 4. The recommended research directions are focused on the three research questions of this dissertation.

Table 4. Recommendations for future research

Research focus	Future research
Data and human-factors	• Human-focused implementation research o Investigate how users interact with explainable AI (i.e., human-machine interface) and how trust affects decision-making o Study the effect of educating physicians on the safe use and adoption of AI tools o Keep involving clinicians and other stakeholders before, during, and after implementation • Leveraging EHR data for better prediction making o Move away from manual labeling practices and investigate the potential of large language models to assist in automated labeling o Focus on data quality and assess potential biases in data
Validity and fairness	• Ensure local validity before implementing in clinical practice o Make local data quality assessment standard practice in combination with local validations o Investigate membership and federated learning models to go from local models to locally adapted models from one overarching model • Address fairness issues with a focus on underlying data biases o Investigate simplified fairness measures and frameworks further to reduce health disparities in alignment with regulatory efforts o Assess fairness for patients that belong to multiple minority groups or are for example mixed-race
Effectiveness	• Focus on external validations and clinical trials to prove local validity and clinical impact o Ensure effective clinical trial designs to determine the effect on patient outcomes, implementation outcomes, and costs o Perform cost-effectiveness and HTA studies o Monitor post-market performance and potential adverse events

9.5. Conclusion

To conclude, the potential of AI in predicting perioperative and critical care patient outcomes is gradually being realized by addressing human and data-related challenges and implementing rigorous model validation and updating processes. Following these steps, the AI tool PERISCOPE was developed to assist clinicians in managing postoperative infections. PERISCOPE demonstrated high predictive performance after locally updating the AI-based prediction model and performed on par with experienced clinicians. The evidence for the clinical validity of AI-based prediction models in perioperative and critical remains sparse, and future studies should focus on evaluating the real-life impact on implementation outcomes, patient outcomes, and costs. To ensure that AI-based decision-making does not increase health disparities, fairness issues regarding subgroup performance should be evaluated per use case. By prioritizing data and

label quality, local validity, and implementation strategies, AI tools could transform clinical decision-making and improve patient care globally.

REFERENCES

1. Ghassemi, M., L. Oakden-Rayner, and A.L. Beam, The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health, 2021. 3(11): p. e745-e750.
2. van de Sande, D., et al., To warrant clinical adoption AI models require a multi-faceted implementation evaluation. NPJ Digit Med, 2024. 7(1): p. 58.
3. Mennella, C., et al., Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. Heliyon, 2024. 10(4): p. e26297.
4. Mandl, K.D., D. Gottlieb, and J.C. Mandel, Integration of AI in healthcare requires an interoperable digital data ecosystem. Nat Med, 2024. 30(3): p. 631-634.
5. Sharma, M., et al., Artificial Intelligence Applications in Health Care Practice: Scoping Review. J Med Internet Res, 2022. 24(10): p. e40238.
6. Knop, M., et al., Human Factors and Technological Characteristics Influencing the Interaction of Medical Professionals With Artificial Intelligence-Enabled Clinical Decision Support Systems: Literature Review. JMIR Hum Factors, 2022. 9(1): p. e28639.
7. McLennan, S., A. Fiske, and L.A. Celi, Building a house without foundations? A 24-country qualitative interview study on artificial intelligence in intensive care medicine. BMJ Health Care Inform, 2024. 31(1).
8. Pinsky, M.R., et al., Use of artificial intelligence in critical care: opportunities and obstacles. Crit Care, 2024. 28(1): p. 113.
9. Zhang, J. and Z.M. Zhang, Ethics and governance of trustworthy medical artificial intelligence. BMC Med Inform Decis Mak, 2023. 23(1): p. 7.
10. Kim, M., et al., Requirements for Trustworthy Artificial Intelligence and its Application in Healthcare. Healthc Inform Res, 2023. 29(4): p. 315-322.
11. Graziani, M., et al., A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. Artif Intell Rev, 2023. 56(4): p. 3473-3504.
12. Linardatos, P., V. Papastefanopoulos, and S. Kotsiantis, Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy (Basel), 2020. 23(1).
13. Chou, Y.L., et al., Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. Information Fusion, 2022. 81: p. 59-83.
14. Rudin, C., Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nat Mach Intell, 2019. 1(5): p. 206-215.
15. Rajput, D., W.J. Wang, and C.C. Chen, Evaluation of a decided sample size in machine learning applications. BMC Bioinformatics, 2023. 24(1): p. 48.
16. van Smeden, M., et al., Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease. Eur Heart J, 2022. 43(31): p. 2921-2930.
17. Hanson, B., et al., Garbage in, garbage out: mitigating risks and maximizing benefits of AI in research. Nature, 2023. 623(7985): p. 28-31.
18. Kilkenny, M.F. and K.M. Robinson, Data quality: "Garbage in - garbage out". Health Inf Manag, 2018. 47(3): p. 103-105.
19. Schwabe, D., et al., The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. NPJ Digit Med, 2024. 7(1): p. 203.
20. Balch, J.A., et al., Machine Learning-Enabled Clinical Information Systems Using Fast Healthcare Interoperability Resources Data Standards: Scoping Review. JMIR Med Inform, 2023. 11: p. e48297.
21. International, H.L., HL7 International FHIR V5.0.0. 2024.
22. Yoon, D., et al., Redefining Health Care Data Interoperability: Empirical Exploration of Large Language Models in Information Exchange. J Med Internet Res, 2024. 26: p. e56614.
23. Saarinen, I.H., et al., Creating an inexpensive hospital-wide surgical complication register for performance monitoring: a cohort study. BMJ Open Qual, 2022. 11(3).
24. Ubbink, D.T., et al., Registration of surgical adverse outcomes: a reliability study in a university hospital. BMJ Open, 2012. 2(3).
25. Veen, E.J., et al., The accuracy of complications documented in a prospective complication registry. J Surg Res, 2012. 173(1): p. 54-9.
26. Kim, D.W., et al., Inconsistency in the use of the term "validation" in studies reporting the performance of deep learning algorithms in providing diagnosis from medical imaging. PLoS One, 2020. 15(9): p. e0238908.
27. de Hond, A.A.H., et al., Perspectives on validation of clinical predictive algorithms. NPJ Digit Med, 2023. 6(1): p. 86.

28. Collins, G.S., et al., Evaluation of clinical prediction models (part 1): from development to external validation. BMJ, 2024. 384: p. e074819.
29. Steyerberg, E.W. and F.E. Harrell, Jr., Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol, 2016. 69: p. 245-7.
30. Van Calster, B., et al., There is no such thing as a validated prediction model. BMC Med, 2023. 21(1): p. 70.
31. la Roi-Teeuw, H.M., et al., Don't be misled: 3 misconceptions about external validation of clinical prediction models. J Clin Epidemiol, 2024. 172: p. 111387.
32. Wong, A., et al., External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. JAMA Intern Med, 2021. 181(8): p. 1065-1070.
33. Sahiner, B., et al., Data drift in medical machine learning: implications and potential remedies. Br J Radiol, 2023. 96(1150): p. 20220878.
34. Xu, J., et al., Federated Learning for Healthcare Informatics. J Healthc Inform Res, 2021. 5(1): p. 1-19.
35. Steyerberg, E.W., et al., Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: An overview and illustration. Stat Med, 2019. 38(22): p. 4290-4309.
36. Makhlouf, K., S. Zhioua, and C. Palamidessi, Machine learning fairness notions: Bridging the gap with real-world applications. Information Processing & Management, 2021. 58(5).
37. NJI. Toeslagenaffaire. [cited 2024 October 25]; Available from: nji.nl/toeslagenaffaire.
38. Balayn, A., C. Lofi, and G.J. Houben, Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. Vldb Journal, 2021. 30(5): p. 739-768.
39. Paulus, J.K. and D.M. Kent, Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. NPJ Digit Med, 2020. 3: p. 99.
40. Blair, I.V., J.F. Steiner, and E.P. Havranek, Unconscious (implicit) bias and health disparities: where do we go from here? Perm J, 2011. 15(2): p. 71-8.
41. Gichoya, J.W., et al., AI recognition of patient race in medical imaging: a modelling study. Lancet Digit Health, 2022. 4(6): p. e406-e414.
42. Qiao, S., et al., Developing an ethical framework-guided instrument for assessing bias in EHR-based Big Data studies: a research protocol. BMJ Open, 2023. 13(8): p. e070870.
43. Carolan, J.E., et al., Technology-Enabled, Evidence-Driven, and Patient-Centered: The Way Forward for Regulating Software as a Medical Device. JMIR Med Inform, 2022. 10(1): p. e34038.
44. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002, and Regulation (EC) No 1223/2009, and repealing Council Directives 90/385/EEC and 93/42/EEC. OJ L 117, 2017: p. 1-175.
45. Medical Devices. [cited 2024 August 23]; Available from: business.gov.nl/regulation/medical-devices/
46. REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). 2016.
47. Health Insurance Portability and Accountability Act of 1996, Pub L No. 104-191, 110 Stat 1936.
48. Parliament, E., European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206—C9-0146/2021—2021/0106(COD)). 2024.
49. Baines, R., et al., Navigating Medical Device Certification: A Qualitative Exploration of Barriers and Enablers Amongst Innovators, Notified Bodies and Other Stakeholders. Ther Innov Regul Sci, 2023. 57(2): p. 238-250.
50. de Hond, A.A.H., E.W. Steyerberg, and B. van Calster, Interpreting area under the receiver operating characteristic curve. Lancet Digit Health, 2022. 4(12): p. e853-e855.
51. Loftus, T.J., et al., Artificial Intelligence-enabled Decision Support in Surgery: State-of-the-art and Future Directions. Ann Surg, 2023. 278(1): p. 51-58.
52. van de Sande, D., et al., Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. Intensive Care Med, 2021. 47(7): p. 750-760.
53. Balki, I., et al., Sample-Size Determination Methodologies for Machine Learning in Medical Imaging Research: A Systematic Review. Can Assoc Radiol J, 2019. 70(4): p. 344-353.
54. Riley, R.D., et al., Minimum sample size for external validation of a clinical prediction model with a binary outcome. Stat Med, 2021. 40(19): p. 4230-4251.

55. Collins, G.S., et al., TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 2024. 385: p. e078378.
56. Vasey, B., et al., Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med, 2022. 28(5): p. 924-933.
57. de Hond, A.A.H., et al., Predicting Readmission or Death After Discharge From the ICU: External Validation and Retraining of a Machine Learning Model. Crit Care Med, 2023. 51(2): p. 291-300.
58. Badal, K., C.M. Lee, and L.J. Esserman, Guiding principles for the responsible development of artificial intelligence tools for healthcare. Commun Med (Lond), 2023. 3(1): p. 47.
59. Lasko, T.A., E.V. Strobl, and W.W. Stead, Why do probabilistic clinical models fail to transport between sites. NPJ Digit Med, 2024. 7(1): p. 53.
60. Hüllermeier, E. and W. Waegeman, Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Machine Learning, 2021. 110(3): p. 457-506.
61. Nagendran, M., et al., Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ, 2020. 368: p. m689.
62. Esteva, A., et al., A guide to deep learning in healthcare. Nat Med, 2019. 25(1): p. 24-29.
63. Singh, R.P., et al., Current Challenges and Barriers to Real-World Artificial Intelligence Adoption for the Healthcare System, Provider, and the Patient. Transl Vis Sci Technol, 2020. 9(2): p. 45.
64. Ahmed, M.I., et al., A Systematic Review of the Barriers to the Implementation of Artificial Intelligence in Healthcare. Cureus, 2023. 15(10): p. e46454.
65. Chipsoft. ChipSoft en Autoscriber werken via AI aan consult van de toekomst. [cited 2024 August 17]; Available from: chipsoft.com/nl-NL/nieuws/854.
66. Santeon, AI op de IC. [cited 2024 August 17]; Available from: santeon.nl/project/slimmesoftwareopdeic/
67. Proctor, E.K., et al., Implementation research in mental health services: an emerging science with conceptual, methodological, and training challenges. Adm Policy Ment Health, 2009. 36(1): p. 24-34.
68. Zabor, E.C., A.M. Kaizer, and B.P. Hobbs, Randomized Controlled Trials. Chest, 2020. 158(1S): p. S79-S87.
69. Han, R., et al., Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. Lancet Digit Health, 2024. 6(5): p. e367-e373.
70. Shimabukuro, D.W., et al., Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. BMJ Open Respir Res, 2017. 4(1): p. e000234.
71. Wijnberge, M., et al., Effect of a Machine Learning-Derived Early Warning System for Intraoperative Hypotension vs Standard Care on Depth and Duration of Intraoperative Hypotension During Elective Noncardiac Surgery: The HYPE Randomized Clinical Trial. JAMA, 2020. 323(11): p. 1052-1060.
72. Tsoumpa, M., et al., The Use of the Hypotension Prediction Index Integrated in an Algorithm of Goal Directed Hemodynamic Treatment during Moderate and High-Risk Surgery. J Clin Med, 2021. 10(24).
73. Hussey, M.A. and J.P. Hughes, Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials, 2007. 28(2): p. 182-91.
74. Vimalesvaran, K., et al., Assessing the effectiveness of artificial intelligence (AI) in prioritising CT head interpretation: study protocol for a stepped-wedge cluster randomised trial (ACCEPT-AI). BMJ Open, 2024. 14(6): p. e078227.
75. Farah, L., et al., Suitability of the Current Health Technology Assessment of Innovative Artificial Intelligence-Based Medical Devices: Scoping Literature Review. J Med Internet Res, 2024. 26: p. e51514.
76. Arnetz, B.B., et al., Enhancing healthcare efficiency to achieve the Quadruple Aim: an exploratory study. BMC Res Notes, 2020. 13(1): p. 362.
77. Zuhair, V., et al., Exploring the Impact of Artificial Intelligence on Global Health and Enhancing Healthcare in Developing Nations. J Prim Care Community Health, 2024. 15: p. 21501319241245847.
78. Goal 3. IMPALA solution. [cited 2024 August 27]; Available from: www.goal3.org/product.
79. Varghese, C., et al., Artificial intelligence in surgery. Nat Med, 2024. 30(5): p. 1257-1268.