A methodology based on quality gates for certifiable AI in medicine: towards a reliable application of metrics in machine learning

Miriam Elia, Bernhard Bauer

Angaben zur Veröffentlichung / Publication details:

Elia, Miriam, and Bernhard Bauer. 2023. “A methodology based on quality gates for certifiable AI in medicine: towards a reliable application of metrics in machine learning.” In *ICSOFT 2023: Proceedings of the 18th International Conference on Software Technologies, 10-12 July 2023, Rome, Italy*, edited by Hans-Georg Fill, Francisco José Domínguez Mayo, Marten van Sinderen, and Leszek Maciaszek, 486–93. Setúbal: SciTePress. https://doi.org/10.5220/0012121300003538.

A Methodology Based on Quality Gates for Certifiable AI in Medicine: Towards a Reliable Application of Metrics in Machine Learning

Miriam Elia[a] and Bernhard Bauer[b]
Faculty of Applied Computer Science, University of Augsburg, Germany
(miriam.elia, bernhard.bauer)@informatik.uni-augsburg.de

Keywords: Certifiable AI, Quality Management, Machine Learning, Healthcare, Metrics, Deep Learning, Performance Evaluation, Algorithm Auditing.

Abstract: As of now, intelligent technologies experience a rapid growth. For a reliable adoption of those new and powerful systems into day-to-day life, especially with respect to high-risk settings such as medicine, technical means to realize legal requirements correctly, are indispensible. Our proposed methodology comprises an approach to translate such partly more abstract concepts into concrete instructions - it is based on *Quality Gates* along the intelligent system's complete life cycle, which are composed of use-case adapted *Criteria* that need to be addressed with respect to certification. Also, the underlying philosophy regarding stakeholder inclusion, domain embedding and risk analysis is illustrated. In the present paper, the *Quality Gate Metrics* is outlined for the application of machine learning performance metrics focused on binary classification.

1 INTRODUCTION

Thanks to astonishing results, the adoption of AI in medicine is moving more and more into the center of attention. Many requirements for a conscious integration of the new technology, especially regarding high-risk contexts, have been published. Recently, the EU released its AI Act that stands as a legislative guideline (European Commission, 2021). However, technical means to realize these requirements in medicine are yet to be developed and standardized. In addition, ”[t]he healthcare application field introduces requirements and potential pitfalls that are not immediately obvious from the ’general data science’ viewpoint” (Jussi, 2021, 1). Challenges regarding the desired adoption of this complex technology into clinical day-to-day life are partly based on the necessity of comprehensive Machine Learning (ML) knowledge to accurately evaluate the system. The present work is part of our approach towards a generic and customizable methodology - introduced in this paper and based on *Quality Gates* (QG) - that comprises existing research on the development of ML models for Certifiable AI in Medicine into guidelines for developers and auditing offices, while paying special attention to end-user perspectives, the inclusion of domain knowledge and risk analysis. The focus lies on defining general guidelines towards metrics selection for a comprehensive evaluation of the ML model, adapted to the respective medical context. Section 2 explains the current legal situation regarding intelligent medical devices with respect to software quality management and metrics for ML in healthcare. In section 3, our methodology's basic concepts are introduced, while section 4 specializes on the *QG Metrics* and presents guidelines for a reliable selection, adapted to the medical context. Finally, section 5 summarizes the present work and derives open research questions.

2 RELATED WORK

Functional & Safety Standards: Since 2021, an updated version of the Medical Device Regulation (MDR) is in place that guarantees the Conformité Européenne (CE), i.e. conformity with ”[...] EU safety, health and environmental protection requirements, as well as with norms set by the International Organization for Standardization (ISO)” (Ben-Menahem, 2020, 1). ISO does not perform certification activities itself, but provides internationally accepted norms, as *DIN EN ISO 9001:2015-11* for process-oriented quality management systems, or *ISO 13485* for medical devices, for instance. Another important concept is safety, i.e. protecting the user from potentially harm-

[a] https://orcid.org/0000-0001-6253-230X
[b] https://orcid.org/0000-0002-7931-1105

Elia, M. and Bauer, B.
A Methodology Based on Quality Gates for Certifiable AI in Medicine: Towards a Reliable Application of Metrics in Machine Learning.
DOI: 10.5220/0012121300003538
In *Proceedings of the 18th International Conference on Software Technologies (ICSOFT 2023)*, pages 486-493
ISBN: 978-989-758-665-1; ISSN: 2184-2833

ful behavior of the software. Functional requirements are summarized under *IEC 61508*, while *DIN EN IEC 60601* and *DIN EN IEC 62304* specifically focus on medical devices. Moreover, Safety Integrity Levels $0-4$ (SIL), i.e. "[...] classification levels indicating safety requirements in safety-critical systems"(Papadopoulos, 2010, 1) are assigned.

Certification & Medical AI: As of now, the EU AI Act is on everyone's mind, aiming to form the "[...] legislation for a coordinated European approach on the human and ethical implications of AI" (European Commission, 2021, 2). This document defines the foundation of AI-based devices in the EU, its philosophy is summarized in (European Commission, 2020), and discussed in further detail with respect to medicine in (Schneeberger, 2020). Currently, the certification process for high-risk medical devices is conducted by an independent authority, i.e. notified bodies. (Ben-Menahem, 2020, 1-3) However, currently, they are not equipped to implement all incoming demands, which could lead to a scarcity of medical devices in the EU (European Commission, 2023, 2-4). For a comprehensive impact analysis regarding the new MDR regulations for risk classes, clinical evaluation, post-market surveillance and notified bodies, refer to (Niemiec, 2022). Current challenges for AI in healthcare are mainly centered around black box models that are able to perform complex tasks, but whose inner workings are incomprehensible for human stakeholders. This could lead to an incorrect application of developed models in the clinical context, "[...] due to methodological flaws and/or underlying biases" (Roberts, 2021, 1), for instance. In (Muller, 2021) generally applicable principles regarding AI in medicine that could form a solid baseline for technical design decisions, are summarized.

Quality Gates & Metrics: A QG is a concept derived from software quality management, and could be defined as "[...] an objective quality assurance gate, that is, a verification procedure, performed either by independent reviewers or by automated scripts" (Paula F., 2006, 34). Their most basic functioning consists of summarizing important criteria regarding specific outcomes that are generated at different points during the software development life cycle (Flohr, 2008, 245). A means of defining criteria for virtual QGs for manufacturing use cases is presented in (Filz, 2020, 8ff), but could be adapted to medical contents, since they are based on the inclusion of domain knowledge. A thorough and comprehensive understanding of the respectively conveyed information is indispensable for ML performance metrics interpretation, especially in medicine, but not necessarily guaranteed (Hicks, 2022, 1). For instance, a very common metric for classification tasks is the *Receiver Operating Characteristic Area under the Curve* (ROC AUC). It is used as primary evaluation metric in popular bench marking tools hosted e.g. on Grand-challenge.org, like the STOIC[1] challenge for 3D computer tomography classification of COVID-19 infected lungs (Boulogne, 2023), for instance. Their metrics selection is based on (Reinke, 2021), according to which ROC AUC and its prominent opposition *Precision-Recall AUC* (PR AUC) both reflect data imbalance (Reinke, 2021, 43ff.). However, there is an ongoing discussion whether or not ROC AUC reflects imbalanced data sets, which is a very common case in medicine (Davis, 2006; Saito, 2015). Also, published paper and bench-marking tools tend to display disagreement regarding the consistent application of both metrics for an empirical analysis (Ribeiro, 2020; Strodthoff, 2020). This inconsistency enforces the necessity to standardize valid approaches.

3 METHODOLOGY BASED ON QUALITY GATES

Our proposed methodology's main objective is to "make auditing simple", and thus provide concrete instructions for the domain-adapted realization of specific legislative requirements in the context of Certifiable AI in medicine, while respecting different stakeholder's needs and specific design decisions' risks. In the long term, such findings could be adapted in a (partially) automated manner to the complete application's life cycle through adapted frameworks and templates for a comprehensive documentation of design decisions. In general, the conceptual foundation is based on the definition of scientifically substantiated *Criteria* for QGs along the complete life cycle of the intelligent software. To the best of our knowledge, a similar adaptation of QGs and ML-certification in healthcare has not yet been published. Attributed to the variety of different ML methods for different medicinal use cases that compose of different data types and tuning objectives, the concrete realization of *Criteria* should be adapted respectively. Structural similarities from a technical viewpoint between use cases should suffice to generalize applied methods, as in (Strodthoff, 2020, 3) where metrics from multi-label protein discovery were adapted to ECG-classification.

General Structure of Quality Gates: In figure 1 the high-level QG's hierarchy adapted to ML-processes is depicted: *QG Data* ensures a clean and informa-

[1] https://stoic2021.grand-challenge.org/

tive data set that is ready for model training, *QG Software* guarantees overall compliance with software engineering requirements, *QG Model* delivers a transparent algorithm that has been thoroughly assessed, *QG Deployment* assures a seamless rollout, while *QG Maintenance* ensures regular monitoring, which could include physician training in the medical sector Only in combination, the whole *QG4Application* is evaluated.[2]

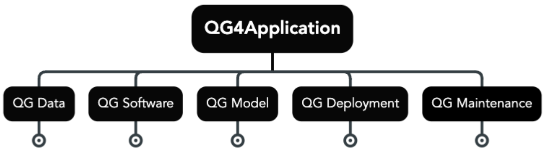

Figure 1: High-Level Quality Gates.

The process-steps depicted in figure 1 represent basic ML-development and should be audited for all levels of risk regarding intelligent applications, refer to (Koshiyama, 2021, 3), where five similar stages of development are defined for ML algorithm auditing in general, or (Oala, 2021, 2) where the authors present a concept for healthcare-specific algorithm auditing.

3.1 Basic Concepts

The following definitions comprise our presented methodology's conceptual foundation, and are partly derived from traditional software quality assessment. In (Koshiyama, 2021), the authors propose *Explainability*, *Robustness*, *Fairness*, and *Privacy* as auditing verticals, which form an important component of Algorithm Audit research (Koshiyama, 2021, 2-3). Our ”auditing verticals” are intended as guidelines of thought when defining *QG Criteria* that should have been analyzed by the responsible party for the respective process step(s). Within our context, fundamental requirements for trustworthy, and thus certifiable AI, such as fairness, privacy, and robustness (European Commission, 2020), are addressed via a profound *Risk Analysis*.

Quality Gate: Significant milestone or decision point during the creation of a ML-based software that, in a body, serve as a quality guideline to assess the software‘s compliance with EU-legislation regarding Certifiable AI in medicine. Project-specific *Criteria* are evaluated against pre-defined desired *Criteria* for the particular use case. Based on the degree of their fulfilment, *Gatekeepers* decide the project's level of compliance, which might lead to re-working some QGs. QGs might be optional or weighted differently regarding their impact. (Flohr, 2008)

Scope: Each QG has access to specific project-based resources or outcomes that are measured by the respective *Gatekeeper*. Its *Scope* includes other QG's outcomes. QGs are arranged in a tree-structure, with growing *Scope* from more project-specific leaf-QGs to more abstract root-QGs. The highest level of *Scope* covers QGs for *Data*, *Software and Model Development*, as well as *Application Deployment and Maintenance*.

Criteria: Basis for QG-evaluation by the *Gatekeeper*. Concrete and use case specific requirements with growing level of abstractness following the *Scope* from leaf to root. Should be adapted to the specific use case if necessary. (Flohr, 2008)

Gatekeeper: Measures the fulfillment of each QG regarding its decision *Scope* depending on the point in time of the application life cycle. It compares predefined, desired *Criteria* with the actual project's outcomes and decides to what extend the system is in compliance. (Flohr, 2008)

Scoring System: It comprises multiple indexes, with the *Compliance Index* as central part, i.e. the ”main index” that comprises the complete application's evaluation in a single number. Its calculation comprises other indexes and the *Gatekeeper's* results. For instance, *QG Data* could be evaluated as stand-alone or embedded within the application's assessment.

Explainability: Since XAI has become a very popular field of research, its inclusion during the ML-based application's life cycle is addressed separately: we follow a similar philosophy, as in (European Commission, 2020), where XAI contributes to the requirement *Transparency*, through providing a pool of methods, that help to ”[...] explain both the technical processes of the AI system and the reasoning behind the decisions or predictions that the AI system makes” (European Commission, 2020, 14). Thus, depending on the respective process step, the application of XAI can have various forms and objectives in support of realizing other *Criteria*, rather than being the center of an evaluation: a developer might use LIME to assess the model's performance regarding learnt features (Ribeiro, 2016) to achieve robustness, while a physician requires a humanly readable explanation.

3.2 Quality Gates in the Application's Life Cycle

Following ISO guidelines, the aforementioned high-level QGs are illustrated as processes during the software's life cycle in figure 2: *Data Management* and all its sub processes, followed by *Model Develop-*

[2] Our proposed methodology focuses on the ML part of the complete medical device, thus, Software Engineering-specific information is only mentioned marginally.

ment and *Software Engineering*, and is concluded with the *Application's Deployment*, and *Maintenance* in the real world. Our suggested methodology aims to assist development teams in form of the inclusion of *Domain Knowledge*, or communication of *QG Inter-Dependencies* based on previous/following QG outcomes during the application's life cycle with the objective to design the application in a way that *Stakeholder's* needs are fulfilled and possible *Risks* mitigated. Also, the auditor will find scientifically grounded guidelines for ML quality assessment tailored to specific groups of medical use cases thanks to our methodology's customizability.

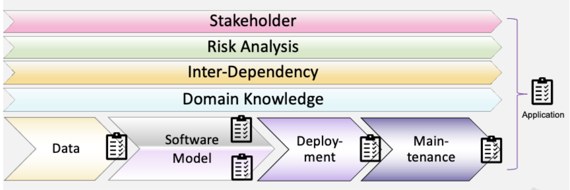

Figure 2: Quality Gates during the ML Life Cycle.

Regarding our proposed methodology's application, it should be considered, that "[a]lthough these stages appear static and self-containing, in practice they interact in a dynamic fashion, not following a linear progression but a series of loops [...]"(Koshiyama, 2021, 3). This may include multiple iterations of the presented processes, starting during development and before certification, while continuing afterwards. For auditing, only a static version of a particular model with its respective data can be assessed, and each modification or re-training automatically requires a new audit[3].

3.2.1 Overall Guidelines

Generally, we defined four guidelines as necessary lines of thought for the definition of *QG-Criteria*, with the objective to unify common ML-concepts in support of different stakeholders involved during the ML-based software's life cycle. Thus, regarding ML for medicine, special focus is placed on the inclusion of *Stakeholders*, while paying special attention to a thorough impact *Risk Analysis* in the real world. Also benefits of the inclusion of *Domain Knowledge* are evaluated, and *Inter-Dependencies* regarding different QGs' outcomes considered.

Stakeholder: Considering all *Stakeholder* needs from a technical point of view enhances the implementation of a successful application that fulfills its intended purpose. The most obvious include *Developer* and *Domain Experts* from a medical background who participate in the software's development, the *Auditor* who is responsible to ensure product compliance with legislation, and finally the *User*, i.e. medicinal personnel and patients, who will one day work with the system. Thus, interdisciplinary teams are advised to be considered standard during the complete development process.

Risk Analysis: The application-wide *Risk Analysis* is realized by the mapping of conducted analysis with differing and specific aims into indexes. Examples include uncertainty estimation, fairness, privacy, transparency, robustness and sustainability.

Inter-Dependency: This guideline is methodology-specific and refers to communicating result-based recommendations between QGs. An example for an *Inter-QG-Dependency* is the recommendation of adequate metrics based on the *QG Data's* analysis and clinical objective of the project, or the effects of *QG Pesudo- /Anonymization* on included meta data as additional features.

Domain Knowledge: Especially in medicine, the inclusion of domain- and use case-specific knowledge is indispensable for efficiently training and accurately evaluating the ML-model, since AI-based software for healthcare is primarily designed to enhance clinical treatment and patient care. Thus, the inclusion of use-case specific domain knowledge should be considered, when designing *Criteria* for leaf-QGs.

3.2.2 Data

The proposed *QG Data* comprises the processes *Source Selection*, *General Preparation*, and *ML Preparation*, and is illustrated in figure 3. When defining the data set composition, a high distribution in data sources is desired, in favor of the algorithm's quality. Further, raw data needs to be analyzed, e.g. with respect to data type, and possibly collected meta data, as well as cleaned from missing values or errors. Then, especially in healthcare, data pseudo- or anonymization is likely to be required. In a final step, since the data is intended to serve as basis for training a ML algorithm, samples need to be annotated correctly, as well as the resulting label and feature distributions analyzed.

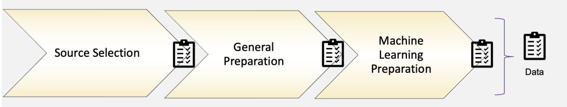

Figure 3: Quality Gates for Data Set Generation.

Bias: Biased data sets propagate their inherent distributions through model training into the application's

[3] For instance, mirroring the *Digital Twin* concept from the industry, a static *Real-World Twin* could be deployed in the application and dynamically updated, while its twin is continuously optimized, including new data.

real world context, which could lead to incorrect and unfair predictions. Thus, the implementation of measurements to detect and reduce bias is important for the assessment of the *Fairness* requirement (European Commission, 2020, 6). Bias analysis is relevant for multiple process steps, especially for *QG Data* and *QG Model*, but also in retrospect during *QG Deployment* and *QG Maintenance*. Within our methodology, such methods are summarized in form of a *Bias Index* for *Risk Analysis*. Other interesting, and promising areas of research that are potentially interesting for ML data set preparation in medicine include synthetic data and multi-modal approaches (MacEachern, 2021).

3.2.3 Model

The presented *QG Model* is divided into four sub processes, *Data Quality and Pre-processing*, *Model Training*, *Evaluation*, and *Validation*, as depicted in figure 4. Depending on design decisions regarding the model's architecture or training objectives, different pre-processing steps should be considered and evaluated against domain-specific requirements. During model training, optimized hyper-parameters are calculated, and different architectures compared on train and validation set, while considering data set specific information from the previous *QG Data*, like class imbalance, for instance. The final test set is applied for evaluating the algorithm's generalization performance by means of domain-embedded and meaningfully interpreted metrics.

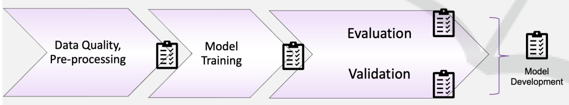

Figure 4: Quality Gates during Model Development.

The extent to which this mostly development-specific information is relevant to present a comprehensive view of the model's behavior for auditing is yet to be defined. However, information included for *QG Validation* like XAI methods that could help to evaluate the model from a different viewpoint, by uncovering wrong patterns learnt or to test the model's robustness through adversarial analysis, for instance (Ribeiro, 2016), are important information to assess the system's overall performance.

3.2.4 Deployment

Other important steps during the application's life cycle, are its deployment and maintenance in the real world. Regarding healthcare-specific requirements however, in our proposed methodology those two inter-related processes are regarded separately. *QG Deployment* is divided into three sub processes *On-boarding*, *Reporting* and *Feedback*, as illustrated in figure 5. Especially in a clinical setting, a close cooperation between the human user, and the intelligent system is evident, which requires a thorough on-boarding phase to support a conscious utilization of the ML-based device. For instance, an appropriate XAI-method could be integrated to analyze model predictions from an additional perspective, but whose interpretation might need further explanations to be humanly interpretable. Refer to (Henry, 2022) for an analysis of physician and intelligent system cooperation.

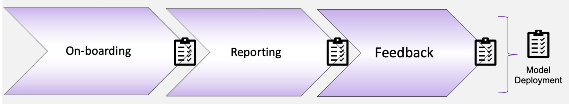

Figure 5: Quality Gates for Deployment.

A valid approach towards achieving a conscious application is educating medical personnel about AI's benefits and risks, as well as necessary basic ML knowledge, depending on their respective degree of interest, i.e. only application or also development of such systems. Other important considerations for this phase include approaches on monitoring the model's behavior in the real world, as well as concepts to transmit and integrate user feedback.

3.2.5 Maintenance

The outlined *QG Maintenance* is divided into four sub processes, *Support*, *Monitoring*, *Optimization*, and *Decommissioning* or *New Data*, as depicted in figure 6. This phase is closely related to the previously defined *QG Deployment*, and partly continues relevant processes. Besides offering user support and training as necessary, as well as repeatedly monitoring the model's real-world performance, algorithm optimization is another important process that should be pursued in parallel, while optionally including new data.

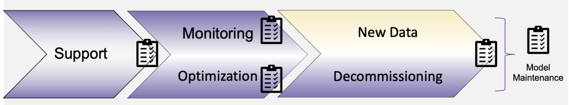

Figure 6: Quality Gates during Maintenance.

These steps are repeated until the application's decommissioning, which should be organized in detail to assure a smooth continuation in the respective real-world setting. The intended tendency is to closely observe the model after its first deployment, and continuously re-assess, but with an increasing distance between iterations, in alignment with MDR regulations (Ben-Menahem, 2020, 3).

4 QUALITY GATE: METRICS

For a reliable model training and performance evaluation within its application context, adequate metrics need to be selected, since their interpretation is not comparable for different tasks with varying objectives and data distributions (Strodthoff, 2020, 6). *QG Metrics* is also relevant after deployment, for maintenance, and to measure clinical success. These multiple points of reference for *QG Metrics* require a thorough documentation of all relevant decisions, their interpretation should be domain-embedded and translated for multiple stakeholders, at best from early development on. However, its results are to be interpreted in combination with *Indexes* relevant for *Risk Analysis*. This section first introduces a general overview of machine learning metrics for the most common binary classification problems, refer to (Müller, 2022) for metrics in image segmentation. In a next step, healthcare-specific challenges that affect metrics interpretation are highlighted. Finally, solid guidelines for a reliable definition of *Criteria* for *QG Metrics* in the broader picture of certifiable AI in medicine are outlined, while addressing domain-specific challenges.

4.1 Metrics for Healthcare

In general, a combination of metrics is necessary for a comprehensive view on the model's performance - no single metric reflects on all desired capabilities (Kelly, 2019, 3). Thus, preliminary material displaying the model's prediction versus true labels for instance, could provide necessary insights to better evaluate the model, for reasonable prediction thresholds, and to enhance the global understanding of applied performance metrics. Regarding metrics as a body, two foundational perspectives could be defined to evaluate model performance depending on the accessible artifacts. Another perspective on ML evaluation that is not part of this research comprises statistical analysis of the model's architecture and its components, as in (Martin, 2021).

Classification: Following this strategy, the model's performance on different classes is measured based on the confusion matrix (Jussi, 2021, 5). Further, this approach is based on *Thresholding* to categorize predictions in either *true* or *false*. Their applicability, including the most widely used ones such as *Accuracy*[4], *Recall*, *Specificity*, *Precision* or *F1-score*, as well as common pitfalls regarding incorrect performance measuring in current research, has been thoroughly studied in (Hicks, 2022).

Ranking: The Ranking-perspective is based on the real valued function learned by the model that returns its confidence, thus accessing the model is necessary. Further, this approach is a threshold-independent performance measurement, includes metrics like ROC AUC (Zhang, 2014, 1822), and could also be referred to for threshold optimization.

4.2 Guidelines for Metrics Application

As a general rule, each selected metric should be accompanied by *Additional Material* that exactly documents its interpretation within the real-world medical context, since measuring clinical efficacy is not trivial (Kelly, 2019, 3). For this purpose, a comprehensive understanding of the underlying data is indispensable, as these represent the link to reality. However, clinical data usually is distributed, heterogeneous and high-dimensional, and multiple sources might first need to be fused following some medical reasoning to represent meaningful input for the ML model (Muller, 2021, 120), which can be an elaborate process. Thus, "[t]he development of quality recommendations and standards for training data sets has to be a community-driven effort of many diverse stakeholders" (Muller, 2021, 120), since high-quality data sets play a crucial factor regarding model performance.

Standards: Some metrics are not symmetric, i.e. the definition which class is positive 1 or negative 0 impacts their outcome and is not interchangeable (Hicks, 2022, 3). A standardized definition marking the disease as the positive class, while healthy samples are defined as the negative class is reasonable for binary classification in healthcare, as proposed in (Jussi, 2021, 9) for instance. Also, the inconsistency regarding metrics selection in general should be addressed by defining a standardized metrics collection for auditing different ML use cases, like e.g. image classification, in addition to "[p]eer-reviewed randomised controlled trials as an evidence gold standard" (Kelly, 2019, 2f.) that accurately measure possible risks and clinical success. The need for further standardization is becoming more prevalent with respect to auditing, and could be realized by the standardized inclusion of a certain metrics combination within popular implementation frameworks.

Bench-Marking: For classifier comparison, benchmarking trained models within different areas of medicine are important to establish a generally accepted performance base line. However, careful consideration is necessary, since some metrics behave differently, depending on the data collection process

[4] Accuracy is a poor measure for imbalanced data sets, and should be replaced by *Balanced Accuracy* (Jussi, 2021, 5-7).

and its resulting diversity (Jussi, 2021, 5). For official bench-marking, either independent real-world test sets that are publicly unavailable should be created (Kelly, 2019, 3), or, another approach are simulation studies based on synthetic data (Friedrich, 2022, 3). Additionally, platforms that provide the necessary infrastructure are required.

Imbalanced Data: As mentioned, imbalanced data is a very common case for medical data. Thus, stakeholders are expected to be aware of this situation and select and interpret metrics accordingly. In contrast to sensitivity and specificity, *positive and negative predictive value* (PPV/NPV) are "[...] influenced by the ratio of disease and healthy cases that happen to be in the test set" (Jussi, 2021, 5), for instance.[5]

Metrics Calculation: Careful considerations are obligatory while designing training and validation versus test data sets, since they are required to be independent for a bias-reduced evaluation and stratified for class-imbalance. Other popular methods for the data pipeline setup during the development process, like cross validation and bootstrapping, are discussed in (Jussi, 2021, 9-12) in great detail, referencing important settings that might need to be audited differently for certification.

Performance Optimization: To select optimal hyperparameters, it is crucial to optimize with respect to the same error measure for comparison (Jussi, 2021, 13), which should be embedded and understood within its medical application area. From a developer perspective, the metric that will be monitored during training for methods like *early stopping* or *learning rate reduction* should be defined carefully.

Domain Embedding: Domain embedded evaluation approaches are expected to be the most resourceful approaches and should be considered as standard for medicine, since the intelligent application's real performance is to be measured and understood regarding its real-world impact (Kelly, 2019, 3). Luckily, "[m]any fields of biomedicine have published their own guidelines on how to evaluate machine learning algorithms [...]" (Jussi, 2021, 9).

Generalizability: Due a high variation in clinical data, achieving a reliable generalizability is challenging but important. A possible solution could include on-site model training to sharpen a pre-trained model towards its specific application context. Further, clinical assessment requires independent and diverse test sets that are capable to measure such abstracts concepts, see *Risk Analysis*. (Kelly, 2019, 4)

5 CONCLUSION

The present work outlines an approach to translate legislation regarding medical AI applications into concrete technical guidelines illustrated for metrics in healthcare. First, the basic concept comprising our proposed methodology is explained in detail, as well as the current situation regarding certification and software quality management for medical AI. Likewise, the philosophy underlying our methodology is outlined while paying special attention to all stakeholders from the beginning, is highlighted. Also, auditing should include all stakeholders' perspectives: the ML-developers', health experts' and/or patients' view of the intelligent application. Further, current ambiguities regarding metrics selection that demand for auditing in medicine to create/retrace commonly accepted concepts to their origins for repeated (re-)evaluation, are addressed. Finally, guidelines for *Criteria* definition(s) that comprise *QG Metrics* are proposed. As of now, we are working on a project in the ECG domain for multi-label classification that will be published as a use case for the proposed approach towards a reliable metrics application.

Our suggested methodology is one possible approach to realize algorithm auditing, and current research should continue to develop standardized compilations for specific ML use cases in favor of the auditing process. Thanks to the multitude and diversity of such use cases, this is not a trivial approach, and the present paper ventures a first attempt to design a comprehensive methodology, presented in more detail for a reasonable selection process for ML performance metrics. To address all existing medical use cases, extensive further research is required, possibly following a mixture of newly proposed technical guidelines, as in (Oala, 2021; Jussi, 2021).

Another principal question that should be further analyzed is to what extend open-sourcing should be made obligatory, since a monopoly on such powerful technologies is questionable. An important part of the outlined methodology includes indexes for *Risk Analysis* that are designed to evaluate more abstract but indispensible concepts such as *transparency* or *robustness*. Further research should consider additional indexes that contribute to a more sound *vue d'ensemble* of the whole model's performance and compliance with legislation, as well as develop technical realizations for relevant points during the software life cycle. While developing such concepts, it might be future-oriented to consider their generalizability towards bench-marking different artifacts like data sets or models, which the presented *Scoring System* might be suitable for. Another crucial aspect, that

[5] PPV is equal to precision for binary classification (Jussi, 2021, 6).

could be included in standard auditing of intelligent medical devices, is measuring metrics or other components via statistical tools such as standard deviation and confidence intervals, refer to (Jussi, 2021) for more information.

ACKNOWLEDGEMENTS

This work was partially funded by the German Federal Ministry of Education and Research (BMBF) under reference number 031L9196B.

REFERENCES

Ben-Menahem, S. M., e. a. (2020). How the new european regulation on medical devices will affect innovation. *Nature Biomedical Engineering*, 4(6):585–590.

Boulogne, L. H., e. a. (2023). The stoic2021 covid-19 ai challenge: Applying reusable training methodologies to private data. Manuscript submitted for publication.

Davis, J., e. a. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML ’06, page 233–240, New York, NY, USA. Association for Computing Machinery.

European Commission, D.-G. f. C. N. C. . T. (2020). *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment*. Publications Office.

European Commission, D.-G. f. C. N. C. . T. (2021). Regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts.

European Commission, D.-G. f. H. . F. S. (2023). Proposal for a regulation of the european parliament and of the council amending regulations (eu) 2017/745 and (eu) 2017/746 as regards the transitional provisions for certain medical devices and in vitro diagnostic medical devices (text with eea relevance.).

Filz, M., e. a. (2020). Virtual quality gates in manufacturing systems: Framework, implementation and potential. *Journal of Manufacturing and Materials Processing*, 4.

Flohr, T. (2008). Defining suitable criteria for quality gates. In *Software Process and Product Measurement*, pages 245–256, Berlin, Heidelberg. Springer Berlin Heidelberg.

Friedrich, S., e. a. (2022). On the role of benchmarking data sets and simulations in method comparison studies.

Henry, K. E., e. a. (2022). Human-machine teaming is key to AI adoption: clinicians’ experiences with a deployed machine learning system. *NPJ Digit Med*, 5(1):97.

Hicks, S. A., e. a. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, 12(1):5979.

Jussi, T., e. a. (2021). Evaluation of machine learning algorithms for health and wellness applications: A tutorial. *Computers in Biology and Medicine*, 132:104324.

Kelly, C. J., e. a. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1):195.

Koshiyama, A., e. a. (2021). Towards algorithm auditing: A survey on managing legal, ethical and technological risks of ai, ml and associated algorithms. *SSRN Electronic Journal*.

MacEachern, S. J., e. a. (2021). Machine learning for precision medicine. *Genome*, 64(4):416–425.

Martin, C. H, e. a. (2021). Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, 12(1):4122.

Müller, D., e. a. (2022). Towards a guideline for evaluation metrics in medical image segmentation. *BMC Research Notes*, 15(1):210.

Muller, H., e. a. (2021). The ten commandments of ethical medical ai. *Computer*, 54(07):119–123.

Niemiec, E. (2022). Will the EU medical device regulation help to improve the safety and performance of medical AI devices? *Digit Health*, 8:20552076221089079.

Oala, L., e. a. (2021). Machine learning for health: Algorithm auditing & quality control. *J Med Syst*, 45(12):105.

Papadopoulos, Y., e. a. (2010). Automatic allocation of safety integrity levels. pages 7–10.

Paula F., W. P. (2006). Quality gates in use-case driven development. In *Proceedings of the 2006 International Workshop on Software Quality*, WoSQ ’06, page 33–38, New York, NY, USA. Association for Computing Machinery.

Reinke, A., e. a. (2021). Common limitations of image processing metrics: A picture story.

Ribeiro, A. H., e. a. (2020). Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications*, 11(1).

Ribeiro, M. T., e. a. (2016). ”why should i trust you?”: Explaining the predictions of any classifier.

Roberts, M., e. a. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 3(3):199–217.

Saito, T., e. a. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):1–21.

Schneeberger, D., e. a. (2020). The european legal framework for medical ai. In *Machine Learning and Knowledge Extraction*, pages 209–226, Cham. Springer International Publishing.

Strodthoff, N., e. a. (2020). Deep learning for ecg analysis: Benchmarks and insights from ptb-xl.

Zhang, M., e. a. (2014). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837.