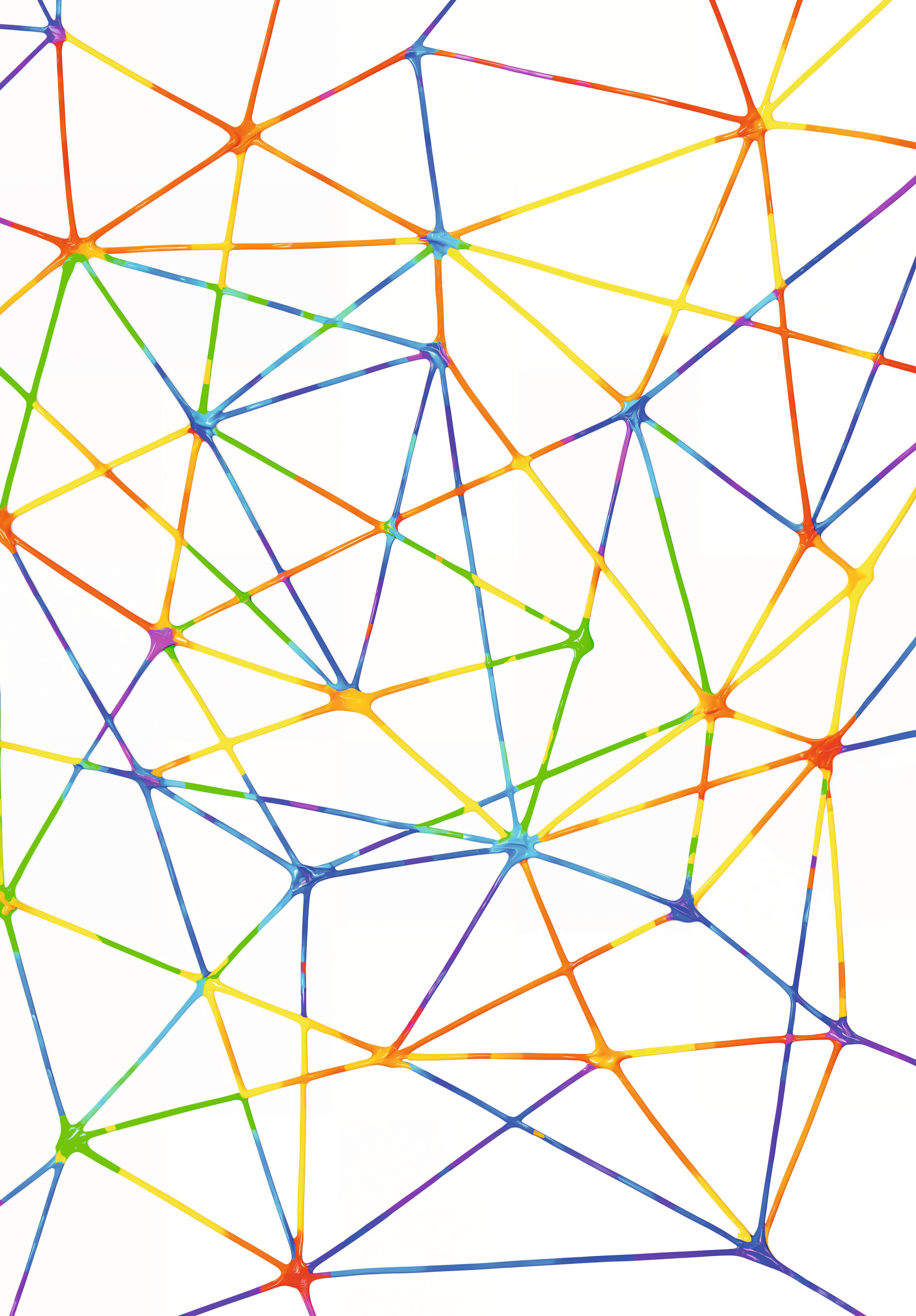
From code to clinic: theory and practice for artificial intelligence prediction algorithms

Hond, A.A.H. de

PART I

PERSPECTIVES ON METHODS FOR CLINICAL ARTIFICIAL INTELLIGENCE PREDICTION ALGORITHMS



2

Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review

Anne A. H. de Hond*, Artuur M. Leeuwenberg*, Lotty Hooft, Ilse M. J. Kant, Steven W. J. Nijman, Hendrikus J. A. van Os, Jiska J. Aardoom, Thomas P. A. Debray, Ewoud Schuit, Maarten van Smeden, Johannes B. Reitsma, Ewout W. Steyerberg, Niels H. Chavannes*, and Karel G. M. Moons*

These authors contributed equally

2.1 ABSTRACT

While the opportunities of ML and AI in healthcare are promising, the growth of complex data-driven prediction models requires careful quality and applicability assessment before they are applied and disseminated in daily practice. This scoping review aimed to identify actionable guidance for those closely involved in AI-based prediction model (AIPM) development, evaluation and implementation including software engineers, data scientists, and healthcare professionals and to identify potential gaps in this guidance. We performed a scoping review of the relevant literature providing guidance or quality criteria regarding the development, evaluation, and implementation of AIPMs using a comprehensive multi-stage screening strategy. PubMed, Web of Science, and the ACM Digital Library were searched, and AI experts were consulted. Topics were extracted from the identified literature and summarized across the six phases at the core of this review: (1) data preparation, (2) AIPM development, (3) AIPM validation, (4) software development, (5) AIPM impact assessment, and (6) AIPM implementation into daily healthcare practice. From 2,683 unique hits, 72 relevant guidance documents were identified. Substantial guidance was found for data preparation, AIPM development and AIPM validation (phases 1-3), while later phases clearly have received less attention (software development, impact assessment and implementation) in the scientific literature. The six phases of the AIPM development, evaluation and implementation cycle provide a framework for responsible introduction of AI-based prediction models in healthcare. Additional domain and technology specific research may be necessary and more practical experience with implementing AIPMs is needed to support further guidance.

2.2 INTRODUCTION

Prediction models have a prominent role in healthcare research and practice. Diagnostic prediction models make predictions about the current health status of a patient, whereas prognostic prediction models estimate the probability of a health outcome in the future [1, 2]. Methods from the machine learning (ML) domain and its broader field of Artificial Intelligence (AI) have seen a rapid increase in popularity for prediction modeling. While the opportunities of ML and AI in healthcare are promising, the growth of complex data-driven prediction models requires careful quality and applicability assessment to guarantee their performance, safety and usability before they are used and disseminated in practice.

A framework for structured quality assessment across the entire AI-based prediction model (AIPM) development, evaluation and implementation cycle is still missing. Such a framework is needed to ensure safe and responsible application of AIPMs in healthcare. For example, it can provide guidance on the appropriate validation steps needed before implementation to prevent faulty decision making based on overfitted models. The absence of such a framework may have contributed to relatively few models having been implemented to date [3]. We define the term AI-based prediction model (AIPM) as follows: a data-driven model that provides probabilistic patient-level predictions of the current presence or future occurrence of a certain outcome (e.g., a certain patient condition), given certain input (e.g., certain patient characteristics, genetic markers, medical images, or other types of features).

We aimed to identify existing guidelines and quality criteria regarding six predefined phases of the AI-based prediction model development, evaluation and implementation cycle. The six AIPM development phases range from preparation and data collection to implementation in daily healthcare practice (see Box 2.1.) and form the core structure and driver for this review. These phases are based on the predominant phases in clinical prediction model research [4, 5]. We performed a scoping review to outline the most important aspects to consider in each phase, while providing pointers to relevant guidelines and quality criteria in the recent literature, focusing on actionable guidance for those closely involved in the AIPM development, evaluation and implementation cycle (e.g., software engineers, data scientists, but also health professionals). We also aimed to identify gaps in the existing guidance.

Box 2.1 Phases[1] of AI prediction model construction

Phase 1. Preparation, collection, and checking of the data: the preparation, collection and checking of the data to facilitate proper AIPM development (phase 2) and AIPM validation (phase 3).
Phase 2. Development of the AIPM: the modelling of the relation between the predictive input variables (features / predictors) and the health outcome of interest, via a mathematical formula or algorithm.
Phase 3. Validation of the AIPM: the testing (validating) how well the developed AIPM from phase 2 predicts the outcome in individuals whose data were not used during AIPM development (so called external validation data), quantifying the AIPM's predictive performance.
Phase 4. Development of the software application: the development of the software application, containing the programming, design, usage and support of the digital packaging of the AIPM.
Phase 5. Impact assessment of the AIPM with software: the assessment of the impact of the usage of the AIPM and software on daily healthcare practice, patient or individual health outcomes, and healthcare costs.
Phase 6. Implementation and use in daily healthcare practice: the implementation of the AIPM in routine care, including maintenance, post-deployment monitoring, and updating.

[1]These phases are primarily introduced to provide clear structure to the article. In practice the order of these phases may slightly differ.

2.3 METHODS

A multi-stage screening strategy was used for this scoping review driven by the six AIPM development phases (Figure 2.1). We searched for relevant academic literature published from January 2000 up to January 2021 in three online databases containing a variety of medical, technical, ethical, and social science literature: PubMed, Web of Science, and ACM Digital Library. The search strings consisted of a combination of search terms related to: i) guidelines, quality criteria, best practices and reporting standards ii) artificial intelligence, including machine learning and prediction modelling in general and iii) topics relating to one of the six phases of AIPM development (see Box 2.1), such as 'data cleaning' for phase 1 and 'impact assessment' for phase 5. For the complete search strings, see supplementary Table S1.

We used the following inclusion criteria for our review process: i) documents (e.g., reports, articles, or guidelines) primarily aimed at the individuals directly involved with the development, evaluation, and implementation of AIPMs (excluding institution or organization wide guidance) and ii) documents with actionable guidance (e.g., clearly defined recommendations on how to develop AIPMs and implement them into practice). The following exclusion criteria were used: i) guidance limited to one medical domain (e.g., cardiology) without generalizing to other domains, ii) guidance limited to one AI technique (e.g., reinforcement learning) without generalizing to other techniques, iii) guidance aimed at governing institutions, iv) documents published before 2000, v) guidance limited to the prerequisites to develop, validate and implement an AIPM (e.g., documents focusing on the development of data infrastructures or legal and governance frameworks), and vi) documents not written in English.

Two reviewers (AdH and AL) performed title and abstract screening of the documents produced by the online database search. Additional literature was added through manually scrutinizing (snowballing) the reference lists of the identified documents. We also asked a convenience sample of 14 AI experts from academia and industry to provide potentially relevant sources (see supplementary Table S3). These additional search strategies were specifically aimed at identifying grey literature consisting of government, institutional or industry documents and websites. The two reviewers performed a full-text screening on all retained literature (including grey literature). Conflicts regarding the eligibility of documents during the screening process were resolved by consensus in regular sessions between the two reviewers.

For the data extraction, two reviewers (AdH and AL) independently identified keywords from each included document which represented the area on which guidance was provided (e.g., development, parameter tuning). Each keyword was mapped to more central topics pertaining directly to the AIPM development phases (e.g., development and parameter tuning were mapped to AIPM training). When applicable to more than one phase, the keyword was placed in a phase-overarching topic (e.g., algorithmic bias). The mapping was adjusted and fine-tuned repeatedly over the course of data extraction and validated based on the input from three co-authors (IK, SN, and MvS). During a second full-text screening round, all identified guidance was extracted according to the topics, summarized, and placed in the review section corresponding to that phase-specific or phase-overarching topic. In our reporting, we adhered to the PRISMA reporting checklist for scoping reviews.

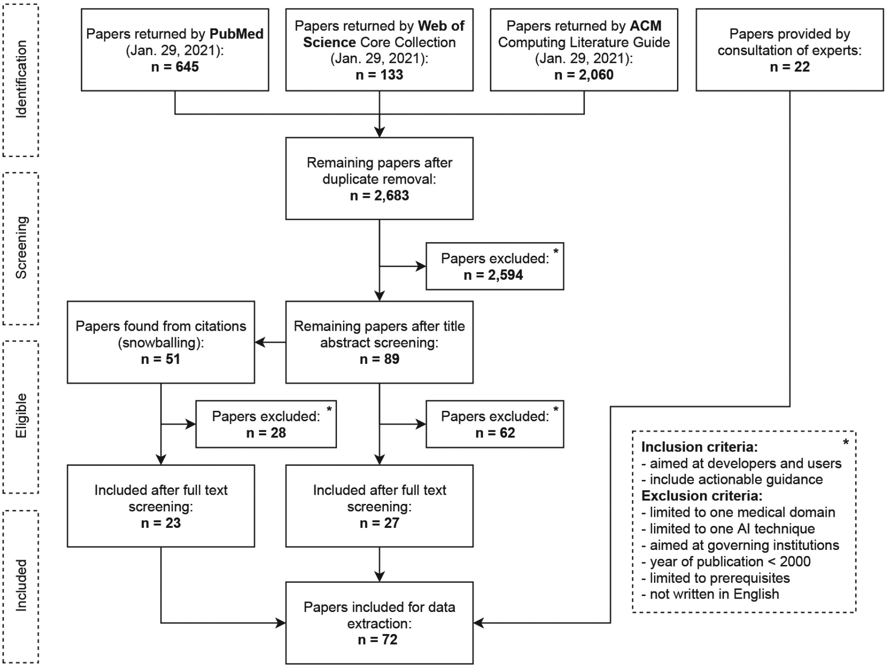

Identification
Screening
Eligible
Included
Papers returned by PubMed (Jan. 29, 2021): n = 645
Papers returned by Web of Science Core Collection (Jan. 29, 2021): n = 133
Papers returned by ACM Computing Literature Guide (Jan. 29, 2021): n = 2,060
Papers provided by consultation of experts: n = 22
Remaining papers after duplicate removal: n = 2,683
Papers excluded: * n = 2,594
Papers found from citations (snowballing): n = 51
Remaining papers after title abstract screening: n = 89
Papers excluded: * n = 28
Papers excluded: * n = 62
Included after full text screening: n = 23
Included after full text screening: n = 27
Papers included for data extraction: n = 72
Inclusion criteria: *
- aimed at developers and users
- include actionable guidance
Exclusion criteria:
- limited to one medical domain
- limited to one AI technique
- aimed at governing institutions
- year of publication < 2000
- limited to prerequisites
- not written in English

Figure 2.1 Multi-stage screening strategy for the scoping review

2.4 RESULTS

After removing duplicates, the search resulted in 2,683 documents. The title and abstract screening reduced this number to 89 documents. Snowballing added 51 documents. A total of 27 papers from online databases, 23 from manual inclusion and 22 from expert consultation, were retained after full-text screening. This led to a total of 72 documents included in the review (Figure 2.1). Data extraction resulted in 138 keywords, which were mapped to 27 phase-specific topics and 6 phase-overarching topics (see supplementary Table S3). In the next sections, the summarized guidance is structured per phase. The phase-overarching topics are summarized in Box 2.2 and further integrated in the phase-specific summaries (as shown in supplementary Table S4).

Box 2.2 Descriptions of identified phase-overarching topics[2]

Algorithmic bias refers to an AIPM that systematically disadvantages individuals belonging to a particular subgroup when there is no a priori medical justification for this discrepancy. Subgroups can for example be based on gender, race, culture, religion, age, sexual orientation, socioeconomic background, ability status and ethnicity. There are two important causes for algorithmic bias: non-representative development data and historical human biases that are reflected in data. The field of AI fairness aims to address algorithmic bias by studying how best to identify and mitigate it.

Transparency and openness entail the possibility to inspect sufficient details on e.g., study design, data selection, analytical scripts, the AIPM model and modelling approach, justifications, and limitations, in a way that could allow others to reproduce the process (e.g., for independent external validation of the AIPM). Recommendations regarding transparency often involve detailed reporting, following relevant reporting guidelines, and sharing of relevant information, code, and data across the different phases.

Interpretability of an AIPM refers to the degree to which a human can understand how an AIPM comes to its predictions or classifications. Being able to interpret an AIPM may facilitate detection of potential errors and biases in its predictions. This may be an important factor in obtaining trust and acceptance by end users (e.g., healthcare professionals and patients). Interpretability and transparency are closely related. For example, an interpretable AIPM may allow a physician to be more transparent about the decision-making process to patients.

Team members, end users, and stakeholders should be considered carefully throughout the AIMP lifecycle (see Box 2.1). It has been recommended that already from the start the AIPM development team must cover a multidisciplinary technical, methodological and medical expertise, consider data and project management, and attend to the diversity of the anticipated end users of the AIPM. Identifying and involving the right expertise and stakeholders in each consecutive phase of the AIPM development, evaluation and implementation cycle is crucial for its success in daily healthcare practice.

Security encompasses the protection of the AIPM and its (personal) data against malicious actors. Two risks particularly concerning an AIPM are the misuse of the (often large amounts of) development and validation data and software vulnerabilities introduced by the new AIPM code and infrastructure. Security measures protecting against these vulnerabilities form part of the AIPM architecture and should be tested before deployment.

Risks refer to any (unintended) consequences of the AIPM's application that threaten the AIPM's safe and effective application. Potential risks are flaws in the design of the AIPM, technical defects, inappropriate or malicious use, process changes, security breaches (see Security above), and disparate outcomes for different use cases or subgroups (see algorithmic bias and fairness above). Safety (for patients and healthcare professionals) should be considered during all phases of AIPM development.

[2] An index on where each phase-overarching topic is further discussed in the article can be found in supplementary Table S4.

2.4.1 Phase 1. Preparation, collection, and checking of the data

Medical problem and context. One of the very first aspects of developing and validating an AIPM as recommended in literature is to clearly specify the medical problem and context that the AIPM will address, and to identify the healthcare setting(s) in which the AIPM is to be deployed [3, 6-15]. Before starting actual AIPM development, it is advocated to first conduct a thorough investigation into the current standard of care, context and workflow [7-9, 11, 13-18], and to provide a clear rationale for why the current approach falls short. For example, via analysis of the needs of targeted end users through observations and interviews, and by involving them from the start in the developmental process [9, 12, 16, 17, 19, 20]. Once a precise (diagnostic or prognostic) prediction task has been formulated, healthcare actions, treatments or interventions should be defined that are to follow from the AIPM predictions [3, 6-11, 17, 21]. Clinical success criteria must be determined and described [3, 9-12, 15, 20, 22], including an analysis of the potential risks of prediction errors [10, 23]. Developers are advised to perform a feasibility check to assess at an early stage whether the expected benefit of the AIPM to the healthcare system outweighs the costs of developing the AIPM, its maintenance, and other consequences of incorrect (or unfair) use of the predictions of the AIPM [8, 9, 12, 15, 22, 24-28].

Patient privacy. The literature advocates that, before starting data collection, the development team should ensure compliance with relevant privacy legislation (e.g., General Data Protection Regulation (GDPR) [29], the Personal Information Protection and Electronic Documents Act (PIPEDA) [30] or the Health Insurance Portability and Accountability Act (HIPAA) [31]) and take measures to protect the privacy of the individuals whose data are used for AIPM development, evaluation, or application [7, 12, 20, 23, 27, 32-36]. Consultation with data protection specialists has been recommended [23]. Legislation may require identification of the right legal basis (such as informed consent) for processing confidential information of individuals [12, 20, 27, 33, 34, 36, 37]. In many cases, individuals must be informed about the processing of their personal data [20, 23, 29, 35, 36, 38]. In the case of using (existing) data that was originally collected for a purpose unrelated to the AIPM (e.g., patient care), there must be an adequate processing basis for re-using these data for AIPM-related purposes [23, 35]. The legal basis can be different for the development and validating versus deployment phases of AIPMs [23, 34]. More specifically, data subjects may not be directly affected by AIPM development but are often affected by AIPM deployment as the AIPM's predictions could influence the treatment decisions of data subjects. Depending on local legislation, it can be required (e.g., under GDPR [29] or the Canadian Privacy Act [39]) to develop a data protection impact assessment [23, 27, 33-35, 40, 41], assign a data protection officer [23, 27, 36], and take measures to conduct data protection oversight, by limiting access only to necessary and qualified personnel [23, 27, 35]. Moreover, taking measures to achieve privacy-by-design [12, 23, 27, 33, 35, 36, 40, 42, 43], such as data minimization [23, 35, 40], encryption [35, 40], or the use of data pseudonymization or anonymization methods is recommended [35, 40]. The use (or absence) of such methods should be clearly motivated [7, 12, 13, 20, 27, 35, 44], especially whenever patient data leave primary care systems [7]. Any trade-offs between predictive performance and privacy should be considered [23]. Finally, under some data protection regulations, individuals have the right to withdraw consent, the right to object, and the right to be forgotten (e.g., under GDPR [29] and the California Consumer Privacy Act [45]), which should be considered and implemented throughout development and deployment stages of the AIPM [12, 23, 36, 40].

Sample size. It is recommended that the amount of collected data is sufficiently large for the intended purpose [7, 10, 12, 14, 20, 22, 27, 46-49], is ideally prespecified [7] and should be clearly reported [3, 13, 37, 48, 50]. The required sample size for AIPM development depends on the specific context, including the used prediction modelling method, the number of features, the proportion of the predicted health outcome (in case of categorical outcomes), and the desired predictive performance [46, 47], which may be linked to a minimal required clinical impact [7]. For regression-based methods [47], and a selection of machine-learning based methods [46], technique-specific a priori sample size calculations are available, although for many model architectures and settings (e.g., semi-supervised learning, decision trees, or convolutional neural networks) no specific guidance was found. If some (closely related) data are already available, it has been suggested to inspect the model's learning curve in that data, setting out prediction performance against the amount of used data, to estimate the required total sample size for a specific use case [46, 51, 52]. For external predictive performance evaluation (discussed in more detail in phase 3), as a rule of thumb, it has been suggested that the sample should at least contain 100 events per outcome [53], but for binary and continuous outcomes more specific sample size calculations are now available [54, 55].

Representativeness. The literature recommends that the collected data are representative of the target population and intended healthcare setting, and sufficiently cover the relevant real-world heterogeneity and diversity [11, 12, 15, 24, 27, 33, 37, 47, 56, 57]. This representativeness criterion is considered crucial to assess and combat algorithmic bias [7, 18-20, 22-24, 26-28, 42, 48, 57-59] and poor calibration [60]. Thorough assessment of the representativeness of the data is strongly advised [6, 10, 11, 13, 18, 27, 37, 48, 56, 57], for which a detailed description of the collected data is required, including the time span of data collection [3, 10-12, 15, 21, 22, 37, 61], the collection site and setting [3, 11, 13, 14, 20-22, 26, 43, 48, 61-63], relevant population characteristics such as gender, age, ethnicity, and relevant medical history [3, 11, 14, 21, 37, 48], and any inclusion or exclusion criteria that were used [3, 6, 10, 11, 13-15, 18, 20, 21, 37, 50, 57, 64, 65]. Finally, revaluation and reporting of any differences between the collected data and the intended target population and setting is emphasized [3, 6, 10, 13, 18, 26, 27, 48, 56, 57], including which groups may be underrepresented in the data with respect to the target population.

Data quality. Extensive assessment of data quality has been widely recommended [6, 10-12, 18, 22, 26, 27, 34, 37, 64, 65]. For both feature variables as well as outcomes, this involves the inspection and description of missing data, consideration of potential errors in measurement, and their underlying mechanisms (e.g., random or systematic) [3, 6, 10, 14, 15, 17, 18, 20, 22, 24, 37, 46, 48, 66, 67]. A clear definition of how and when each variable was measured should be provided [3, 6, 10, 12-15, 17, 21, 22, 28, 37, 48, 50, 58, 62, 64, 65], including specification of measurement instruments or tools (e.g., make and model of devices). Any known data quality risks and limitations should be reported and related to potential impact on the AIPM's predictions and its validation (with special attention to algorithmic bias) [3, 6, 20, 22, 27, 33, 34, 37, 42, 56]. An additional validity check could be performed by randomly sampling a portion of the data and manually checking it for errors [25, 61]. The proportion of errors should be reported [61]. The literature also recommends the installation of a process through which data errors can be corrected [42, 61]. Note that when such a process is installed, it should also be employed during implementation and not just during model development. It must be clearly identified whether data were collected retrospectively or prospectively [10, 13, 14, 21, 48]. Prospective data collection may be preferred as it more closely matches the real-world operating conditions [57]. It was pointed out that one should be aware of potential quality risks of routinely collected data as such data are often collected for a different purpose [57, 68].

The literature places a particular emphasis on the quality of outcome data, more specifically the reference standard or 'ground truth'. A clear rationale on outcome data collection needs to be provided (e.g., an expert panel, biopsy, clinical determination via laboratory tests), and any potential quality issues [3, 6, 10, 13, 14, 21, 48]. In case the outcome data were manually labeled, the AIPM development and validation team are urged to precisely specify how and by whom data were labeled, including the level of experience of the labelers, and elaborate on relevant pitfalls or difficult cases [7, 8, 14, 21, 48, 64-66]. Ideally, to ensure label quality and prevent bias in AIPM evaluation, it was advised that this is a well-defined and controlled process [48, 67], where experts labelling the data work independently from each other [7, 21], and are *not* directly involved in performance assessment of the AIPM [14, 48]. Depending on the exact procedure, inter-observer variability or test reproducibility [7, 14, 21, 48] should be calculated to obtain an assessment of label quality.

Data preprocessing. To prepare data for the consecutive phases, or handle identified data quality issues, data preprocessing steps may be applied. Such preprocessing steps can include splitting the data into different subsets (e.g., train, tuning, and test sets), augmenting data, removing outliers, re-coding or transforming variables, standardization, and imputation of missing data [6, 10, 17, 24, 46, 48, 49, 68]. The literature stresses that detailed description of any pre-processing steps applied to the raw data should be provided, including software used to perform the processing steps [3, 6, 10, 11, 13-15, 22, 50, 61, 62, 64, 65]. Missing data imputation is generally recommended over complete case analysis where incomplete data are excluded, but this should depend on the underlying missing data mechanism (missing completely at random, missing at random, or missing not at random) [6, 17, 46, 49, 68]. Any data augmentation should be carefully considered against the potential introduction of bias, and model developers are advised to collaborate with domain experts on these preprocessing steps [15, 22, 48]. Finally, the literature stresses that data splitting actions, must happen *before* any other preprocessing steps are applied (e.g., missing data imputation or standardization) [24, 69, 70]. This is crucial to prevent information leakage between data subsets, which leads to overoptimistic AIPM predictive performance.

Data coding standards. To facilitate interoperability, and easier adoption of the AIPM into healthcare settings, it has been recommended to align data management with relevant coding standards and widely adopted protocols [20, 27]. Relevant standards may include SNOMED CT for coding clinical data, ICD-10 and OPCS4 for clinical conditions and procedures [20]. Additionally, adopting data exchange protocols in the final AIPM software design has been recommended, but is discussed later in the article (in phase 4, about development of the software application).

2.4.2 Phase 2. Development of the AIPM

Model selection and interpretability. The literature indicates that the following aspects may affect the choice for a certain modelling technique (e.g., regression, decision tree, neural network): prediction performance, interpretability, the familiarity of the modelling technique to the end user, computational requirements, development and validation costs, maintenance, privacy, sample size, and the structure of the data [6, 10, 15, 17, 18, 22, 23, 71]. It is recommended that any motivations for choosing a modelling technique should be clearly articulated [6, 7, 10, 13, 20, 23, 26, 27], including benefits and potential risks associated with the chosen technique [6, 18, 20, 23, 26, 27, 33]. Facilitating interpretability of the AIPM, e.g., by providing insight into the impact of each feature or predictor on the predicted outcome [10, 13, 18, 46, 56, 72, 73], is frequently mentioned as an important aspect for AIPM acceptance into healthcare practice [8, 26, 27, 41, 46, 72]. Important to note is that the term AIPM interpretability - in this scoping review - does not imply causal interpretability (e.g., high feature impact does not imply causal influence of that feature on the actual health outcome). Interpretability may help to detect trivial and erroneous AIPMs [11, 24], provide medical domain experts with a possibility to discuss whether the associations on which the AIPM relies are likely to remain stable [7, 24, 61], help to identify algorithmic bias [11, 22, 24, 26, 41, 42], provide information on where the AIPM could be most easily attacked [24], or how the AIPM may behave under dataset shift [11]. Neural networks are for example recommended for high volume, dense, and complex data types [6, 74], but they are also considered black boxes [23, 26, 34], for which additional model-agnostic interpretation tools (explainable AI) are needed to give insight into the importance of individual features for the predictions [6, 23, 26, 34, 56, 75]. This is in contrast with linear regression and decision trees, which have been considered inherently interpretable approaches. Irrespective of the modelling choice, facilitating interpretability is generally encouraged [6, 23, 26, 33, 34, 40, 41, 56, 62, 71], in particular when AIPMs rely on sensitive social and demographic data, or if the AIPM's predictions significantly affect healthcare decision making and a patient's treatment [18, 22, 40]. Moreover, under the GDPR [29], patients have a right to an explanation that enables them to understand why a particular decision was reached [36, 40, 41]. If a form of interpretability is required, the underlying reasons should be made explicit [15, 41].

Training the AIPM. Training (or fitting) the AIPM is the process of determining the values of any model parameters (e.g., also called weights, or coefficients) of the AIPM. Beside model parameters, AIPM development involves choosing hyper-parameters, which influence model training and design, but are not necessarily part of the AIPM itself (e.g., penalization factors of shrinkage, learning rates, or the depth of tree-based methods). Automatic optimization of hyper-parameters (also referred to as *tuning*) has been recommended [15, 24, 67, 76, 77], for example, via nested cross-validation, or using a small representative held-out tuning data set. To foster transparency and replicability it is advised that any details about training and hyper-parameter optimization procedures should be reported, including the final values of the (hyper-)parameters, the number of intermediate models trained to come to the final model, and an evaluation of predictive performance on the training data [3, 6, 7, 13, 14, 50, 61].

Internal validation. The goal of internal validation is to assess the predictive performance of an AIPM in data that are unseen with respect to model training but come from the same population and setting.

To assess AIPM performance, the literature stresses that data should be strictly separated into training, tuning and test sets [6, 7, 11, 77], possibly stratified by the outcome event [15, 24] to prevent data leakage, which can result in optimistically biased evaluation [6, 11, 24, 69]. Here, the training data is used to train the AIPM, the tuning data for optimizing the hyperparameters, and the test data for assessing the AIPM model performance. Variations on the simplistic 'split sample' validation have been suggested for better data efficiency and heterogeneity assessment (e.g., k-fold cross-validation or bootstrapping). Especially for small datasets, a cross-validated procedure is recommended [6, 24]. The cross-validated procedure should incorporate all processing steps (standardization, imputation etc.) on the data to prevent data leakage [15, 69]. The split of the data and any potential repeats of this splitting procedure should be reported [6, 13, 50].

Following the literature, the performance evaluation should be based on at least discrimination and calibration [5, 6, 10, 15, 17, 49, 57, 78]. Discrimination refers to the ability of the AIPM to distinguish between subjects with and without the outcome of interest. It is recommended to define the metrics used to measure discrimination prior to the validation [6, 7, 10]. The chosen metrics should correspond with the intended medical use and should be chosen in close collaboration with domain experts (e.g., an AIPM estimating risk of breast cancer should be highly sensitive) [7, 11, 13-15, 18, 19, 56, 79, 80]. Discrimination is commonly quantified by the area under the receiver operating characteristic curve [14, 15, 17, 48, 49, 57, 69]. In case of a clearly defined probability threshold, other metrics could also be used like sensitivity (also labeled: 'recall') and specificity, or the positive and negative predictive value (also precision) [8, 15, 19, 72, 80]. Note that fixed probability thresholds are not always considered necessary and when they are, they should be carefully determined in collaboration with medical experts [81].

Calibration refers to the concordance between predicted and observed probabilities. A calibration plot is the recommended method to evaluate calibration [10, 17, 49, 57, 60]. Discrimination and calibration evaluation metrics should be documented for all data sets [6, 13, 18]. It is recommended to calculate confidence intervals to accompany these metrics [7, 8, 13, 14, 21, 22, 24, 26, 48, 61].

For some application types, Decision Curve Analysis (DCA) is considered a valuable addition to the discrimination and calibration of the AIPM. This performance assessment quantifies how the AIPM could impact patient care within the relatable workflow. Unlike discrimination and calibration, DCA derives the clinical utility from the predictive performance [5, 10, 17, 49, 68, 72]. Promising results in a DCA can provide a clear indication that an AIPM could benefit daily healthcare practice. It could therefore serve as a precursor (but not a replacement) of a prospective impact study or more fully developed cost-effectiveness analysis (see phase 5).

Measures to reduce risk of overfitting. If an AIPM is adapted too much to the training data, and therefore its predictions no longer generalize well to new individuals not used for the development of the AIPM, the model is said to be overfitted [7, 46, 57, 60, 76, 78]. Often mentioned factors contributing to overfitting are a small sample size in combination with many candidate features, perfect separation on rare categories, and a large imbalance resulting in a small number of events for one of the outcomes [10, 46, 49, 72, 76, 77, 82]. To prevent overfitting, a multitude of strategies are available, often aimed at reducing AIPM complexity. It has been widely recommended to report any measures taken to prevent overfitting [3, 6, 7, 11, 14]. One commonly referred strategy is feature selection [6, 14, 24, 46, 76], for which it is explicitly recommended that selection should work independently of model training (unlike in methods as forward and backward selection) and is best informed - a priori - by medical expert knowledge or existing literature [6, 17, 76]. Other suggested strategies to combat overfitting are dimensionality reduction [46, 76], which can be implicit (e.g., common in neural networks) [76], and explicit penalization of complexity (e.g., regularization) [17, 49, 76]. It should be noted that when the sample size is simply too small, even penalization methods have been shown ineffective to mitigate overfitting [83, 84].

Measures to identify and prevent algorithmic bias. The literature indicates that tools to identify and mitigate algorithmic bias should also be developed in the AIPM development phase when applicable. First, a definition of fairness should be chosen that corresponds with the AIPM's intended use [18]. This definition should be integrated with model development as part of the AIPM's evaluation metrics [22, 26, 28]. Examples of fairness metrics are outcome parity

[22, 23, 28, 42, 43], true (false) positive (negative) rate parity [22, 23, 28, 42, 43, 79], positive (negative) predictive value parity [22, 42, 43], individual fairness [22], counterfactual fairness [22, 26, 43, 59], and equal calibration [23]. Developers are advised to make the chosen fairness metrics available in a Fairness Position or Bias Impact Statement that is reviewed by stakeholders [22, 23, 27, 28, 62]. They are also advised to avoid modelling techniques for which it is altogether impossible to evaluate algorithmic bias in an AIPM, for example due to the high dimensionality of its architecture [22].

Upon identification, algorithmic bias should be addressed by employing an appropriate mitigation strategy during AIPM development, which may be different for different applications and domains. When the bias is caused by unrepresentative training data, the main recommendation is to redo the data collection to rectify this [7, 18-20, 22-24, 26-28, 42, 48, 57-59]. Unrepresentative training data may also be addressed by undersampling the overrepresented group or oversampling the underrepresented group [23, 43]. However, this may cause miscalibration of the model predictions and should be used with caution [85]. The most popular recommendation addressing other causes of algorithmic bias (e.g., historical human biases reflected in the data) is to exclude or reweigh the features causing the algorithmic bias [22-24, 28, 42], although this may not eliminate the bias altogether. Alternatively, the predictions themselves can be reweighed by adjusting the probability threshold per subgroup [42, 43]. Lesser mentioned recommendations consist of the application of fairness optimization constraints during AIPM training [42, 43] and the development of separate models per specific subgroup [23].

Note that the preconceptions and biases of designers can be replicated in their modelling choices [22]. It is therefore considered important to compose a diverse development team [17, 22, 23, 26, 28], create awareness and involve stakeholders in design choices [22, 24, 26, 27, 72]. Also, developers should keep evaluating algorithmic bias at every stage of the development process [33].

Transparency of the modelling process. The literature advocates that the final AIPM structure should be described in detail, covering input, outputs, and all intermediate layers or parameters [3, 13, 14, 50]. To facilitate transparency and reproducibility of the developmental process, the used computational architecture, high-performance techniques, software packages, and versioning (data, model, configurations and training scripts) should be reported [6, 13, 18, 50, 64, 65, 67]. Code for the complete model building pipeline should be published in well-documented scripts with computer environment requirements when possible [6, 7, 11, 13, 18-20, 24, 26, 28, 34, 50, 62, 64, 65], including statements about any restrictions to access or re-use.

2.4.3 Phase 3. Validation of the AIPM

External performance evaluation. In practice, an AIPM is likely to be applied in a setting that differs from the setting in which the AIPM was developed, which may have an impact on AIPM performance. In contrast to internal validation (phase 2), external validation is the application of an existing model without any modifications to data from a different population or setting compared to model development (see Generalizability below). The literature highly recommends external validation for all AIPM applications when applied to a new setting [3, 15, 17, 49, 86]. Similar to internal validation of the AIPM, external AIPM model validation can be based on discrimination (area under the receiver operating characteristic curve, sensitivity, specificity, positive and negative predictive values), calibration (calibration plot) [5, 6, 10, 17, 49, 57, 78], and Decision Curve Analysis [5, 10, 17, 49, 68, 72]. When possible, the literature recommends the comparison of current best practice (e.g., an existing prediction model or medical decision rule) to the AIPM performance [7, 11, 13, 14, 18].

External validation can be performed on retrospective or prospective data. Although prospective validation is rare, it is preferred by the literature [5, 13, 57], as it provides a better idea of the AIPM's true applicability to medical practice and allows the healthcare professionals to identify and review errors in real time [19, 72]. External validation is ideally performed by independent researchers from other institutions or settings [3, 7, 8, 18, 24, 68, 72]. The necessity for external validation by independent researchers may depend on the risks posed by the application (for example based on the level of autonomy of an AIPM) [80].

Generalizability. Generalizability refers to the AIPM's ability to generalize its performance to a new setting. Poor generalizability may be caused by overfitting (see phase 2) or development data that were unrepresentative for the new setting (see phase 1). The literature recommends to assess generalizability on external data from a different time period, place, or healthcare setting [3, 7, 8, 11, 17, 18, 24, 57, 68, 72, 79].

To ensure the generalizability of the AIPM to the intended healthcare setting, developers are advised to extensively validate the model for representative data from that setting [6-8, 10, 11, 13, 14, 24, 26, 57, 64, 67, 68, 72, 77, 79, 87, 88] (see phase 1, Representativeness). The intended healthcare setting may be different from the population or setting on which the AIPM was originally developed (e.g., an AIPM developed at a tertiary care center applied to a smaller hospital). It is advised that the size of this validation data should follow the available sample size recommendations for AIPM validation (see phase 1) [53-55]. Developers are urged to clearly describe any differences between the development and validation data where possible [13] and report other sources potentially affecting generalizability [7, 10, 24]. Still, AIPM updating, site-specific training or recalibration might be needed to adapt an existing AIPM to a different healthcare setting [3, 5, 15, 60, 68, 72]. Statistical updating methods are available for regression-based models [89, 90]. For AIPMs outside of this context no specific guidance was found.

Performance analysis by population subgroups or specific problematic use cases is recommended to identify algorithmic bias [10, 11, 23, 26, 27, 43, 61, 72, 79, 91]. Note that such an analysis may be limited by small sample sizes of certain subgroups. The literature advises to discuss and explicitly report any identified sources of algorithmic bias, so that end users know for whom the AIPM's predictive performance is subpar [7, 18]. Many systems will display some unfairness in their outcomes, and therefore a baseline comparison with the algorithmic bias of the current systems may be considered [18].

2.4.4 Phase 4. Development of the software application

Interoperability. The ability for AIPMs to interoperate with various existing digital infrastructure of hospitals and clinical care centers is essential for their successful integration into healthcare practice. Following existing standards from the industry was recommended as this supports the interoperability of

AIPMs [15, 16, 20, 27] (e.g., ISO/IEC JTC 1/SC 42 [92] or the IEEE 7000-2021 [93]). This applies to data coding standards as mentioned in phase 1 of this article, but also to data exchange standards (e.g., FHIR [94] and the HL7 framework [95]). Such standards provide (among other aspects) guidance on what data formats to use, how they should be exchanged between system components, and reduce the risk that data are accidentally misinterpreted due to slight differences in meaning of variables (semantic interoperability). For wearable devices, following the ISO/IEEE 11073-10418:2014 [96] standard is advised [20].

Moreover, multiple articles recommend the use of open source or publicly available libraries in the software implementation of the AIPM [20, 27] to increase the accessibility of the AIPM as a whole. The NHS guide to good practice for digital and data-driven health technologies goes as far as to recommend that all new digital health services, including AIPMs, should be made internet-facing from day one (and follow the Representational State Transfer design principles) to promote accessibility and reduce complexity and costs of incorporating them in the digital infrastructure of organizations [20].

Human-AI interaction. A proper design of how end users can interact with the AIPM is crucial for its adoption, and effective and safe use in daily healthcare practice. What constitutes a good design depends on the domain, healthcare setting and intended end users. End users interacting with the AIPM can be healthcare professionals, auditors, or patients (e.g., physicians may need to communicate about the AIPM with patients [16]). Many of the recommendations for human-AI interaction design come from the general human-computer interaction literature and current standards for general medical software design. Recommended standards are ISO 9241-210:2019 [97] for interactive systems and the IEC 62366-1:2015 [98] on application of usability engineering to medical devices [20]. At the software development stage, it has been recommended to include experts in user interface design [7, 16]. Designing a good user interface and interaction requires careful consideration of the cognitive load of the end users [8, 16, 68, 99, 100], by showing only relevant information in the right context, and by allowing adjustment of its behavior by end users [99].

A widely suggested minimum criteria for AIPM user interaction design is that it becomes clear to end users what the AIPM's intended use is [27, 79, 88, 99].

Providing a model facts label should be provided to the end users is advised, including the system's technical specifications, statistical working, limitations, fairness criteria and validation, implementation disclaimer, and links to process logs [22, 101].

To arrive at a good design, repeated extensive user experience testing is recommended [9, 16]. The AIPM should be evaluated according to how it interfaces with the end user, and how well the AIPM and the user perform together in a typical environment [8, 100, 102, 103]. It was proposed that such evaluations can, for example, be done via reader and user studies [8, 102, 103]. Tools such as a system usability scale (SUS) have been suggested as a quick and useful way of capturing user feedback [20].

Careful attention should be paid to inclusiveness and broad usability of the design [20, 22, 27, 62], for example by considering the digital literacy of the end users [20, 22, 27]. Multiple sources state that the design should match social norms, and make sure its presentation does not reinforce stereotypes (e.g., regarding a pre-specified fairness position or bias impact statement, see phase 2) [22, 26, 27, 33, 99].

Moreover, the AIPM should have built-in mechanisms that protect the end user and patient from potential risks to its safe application (e.g., overconfidence in the AIPMs predictions or automation bias). These mechanisms should detect situations beyond the capabilities of the AIPM [8, 99], and share the confidence in the predictions with the user [8, 22, 27, 99]. Additional information may be required explaining how the confidence level relates to the input data [23, 42, 61]. It was recommended to carefully consider whether predictions should be presented in a directive fashion (by also proposing decisions), or in an assistive way (e.g., by only showing estimated probabilities) [15, 22, 41, 68, 86, 88].

The literature advised that the design should facilitate AIPM interpretability (see also Box 2.2. and the section on model selection and interpretability in phase 2) and allow end users to visually see the link between the input data and the predicted output [7, 8, 22, 27, 33, 61, 99] in a comprehensive way [22, 23, 26, 27, 41, 43, 62], and encourage giving feedback, correction and refinement about the AIPM's predictions [99]. Also, the design should enable the patient to request a review of an AIPM supported decision [63], and may need to provide the possibility to delete data (depending on local legislation, see phase 1 on Patient privacy) [12, 23, 36, 40].

Facilitating software updating and monitoring. From a user interaction design perspective, it has been recommended that decisions are deterministic (consistently giving the same output for a certain input) [8], and that updates of or adaptations to the AIPM should happen cautiously [99]. End users should be notified clearly about any changes in the AIPM [27, 99], and AIPM software should have the ability to roll back to previous versions, in case an update results in significant problems [20, 67].

Finally, as monitoring and auditing of AIPMs in practice is widely recommended (covered in more detail in phase 6), the developed software should facilitate this [8, 22, 27, 33, 58, 62, 104]. This means adequate logging and traceability of predictions and decisions is required and the AIPM interface should provide sharing of performance data with end users to enable ongoing monitoring of both individual and aggregated cases, quickly highlighting any significant deviations in performance [8, 27, 61, 67]. Such monitoring options should preferably be customizable by the user [79, 99].

Security. The principles of security and privacy by design mandate built-in data and software protection throughout the AIPM lifecycle [12, 35, 40, 42, 43], which is a central requirement in the GDPR [105]. Cybersecurity standards provide guidance on how to approach this [20, 23, 27], for example ANSI/NEMA NH 1-2019 [106], NEN 7510 [107], MDCG 2019-6 [108], ANSI/CAN/UL 2900-1 [109], Medical Device Cybersecurity Working Group on medical device cybersecurity [110], Food and Drug Administration on cybersecurity [111], ISO/IEC TS 27110:2021 [112], ISO/IEC 27032:2012 [113], ISO/IEC 27014:2013 [114], and ISO/IEC 27002:2013 [115]. This might for example entail an initial risk assessment of vulnerabilities in data and software, including the risk of re-identification [34], the risk of data loss and manipulation [34, 35], and the risk of adversarial attacks [15, 22, 23, 27, 35, 42, 59]. Techniques that make the AIPM more robust to these vulnerabilities can be implemented, like converting data to less identifiable formats [23], adding random noise to the data [23, 32, 40], federated learning [23, 32, 40], saving personal data across different databases [32, 35], and adversarial ML techniques such as model hardening and run-time detection [22, 42, 43, 59]. Code review by an external party and staying up to date on security alerts for code derived from third parties are also recommended [23, 35]. All security measures should be tested before full deployment [79] (also see Software testing). The level of the required security measures will depend on the impact a potential security breach might have on the individuals involved, the type of AI deployed, and the risk management capabilities of the organization [23, 24, 35, 40]. The timeframe within which security updates will become available should be reported [27].

An incident response plan anticipating a potential security breach is recommended before deployment (also part of western legislation [104, 105, 116]), describing how incidents will be addressed and who takes responsibility with relevant contact information [23, 35, 61]. When new software vulnerabilities come to light, they should be documented and reported [33, 61], and so should any changes made to the AIPM in response to an attack after thorough testing [8, 23, 35, 61].

Software testing. AIPM software developers are recommended to follow relevant existing international standards with regard to software testing, such as the IEC 62304:2006 [117], the IEC 82304-1:2016 [118], IEC 62366-1:2015 [98], ISO 14971:2019 [119], Food and Drug Administration principles of software validation [120] , and Food and Drug Administration guidance for off-the-shelf software use in medical devices [121]. Deliberate stress tests like load testing, penetration testing, integration testing and unit testing are important for the verification of the AIPM from a software perspective [8, 27, 35, 48, 67, 79]. Each different context of use may require separate software testing to ensure reproducibility of results across different situations, computational frameworks, and input data [58, 62, 87]. These testing requirements depend on the level of reliability needed and the risks posed by the AIPM in healthcare practice [27]. These types of tests are also recommended to assess the effectiveness of the security measures taken and to detect new security vulnerabilities (see Security). They should be repeated regularly to monitor the data and software security during the AIPM lifecycle [23, 27, 35].

2.4.5 Phase 5. Impact assessment of the AIPM with software

Feasibility study. An impact assessment is performed to determine the clinical benefit of the AIPM for healthcare practice. It is important to note that a good performance of the AIPM in terms of discrimination and calibration (phases 2 and 3) does not necessarily translate to clinical utility [5, 24, 72].

A feasibility study or implementation pilot is recommended preceding an impact study to ensure correct and safe use in healthcare practice [8, 16, 72]. This type of study consists of repeated live clinical tests in which variation is key to understanding the functionality of the technology and workflow [9, 16]. By adhering to the 'plan, do, study, adjust' process, adjustments can be made frequently and rapidly to optimize the workflow [9, 16].

The literature advises to clearly define the intended use and intended users in preparation of both the feasibility and impact study [12, 19, 64, 65]. It is also recommended to report any differences in healthcare setting between the current and previous (validation) studies [68] and to state the inclusion and exclusion criteria at the level of the participants and input data [25, 64, 65]. A description of the integration into the trial setting is highly recommended, including onsite and offsite requirements, version number and other technical specifications [25, 64, 65], but also the human-AI interaction involved (e.g., assistive versus directive, see phase 4) [48, 64] and the patient treatment strategy associated with the AIPM outcomes [64, 65]. It is emphasized that potential interventions included in the patient treatment strategy following from the AIPM decision support should have a solid scientific basis [68]. Stakeholders have preferably given informed approval of the development and clinical application of the AIPM [87].

Risk management. Risk management is highlighted as an important part of the impact assessment, alongside the preparations for a comparative study [25, 42]. The literature recommends the identification of potential sources of risk, extreme situations, and failures before the onset of the study [27, 56, 58]. Determining corresponding safety critical levels and quality checks is advised [27]. Special attention may be paid to accidental misuse and manipulation of the AIPM. Implementers are urged to report errors, failures or near misses occurring during impact assessment and afterwards [26, 27, 42, 61, 64, 65]. A risk management plan can help to execute the monitoring, reporting and mitigation of risks encountered in healthcare practice [12, 18, 20, 25, 27]. This plan can for example describe the roles and responsibilities of the participants [25], the process for assessing and logging potential risks [12, 20, 26, 27, 42, 61], a pathway to report potential risks [12, 26, 27, 42, 62], and the process to address these issues in practice [12, 42, 62]. Some sources suggest that the assessment should be proportionate to the risk posed by the AIPM [27, 42].

Impact study. In terms of the impact study design, a prospective comparative study is recommended [5, 7, 19, 24, 57, 68, 72, 86, 87]. In a comparative study, the effects on clinical outcomes and decision making are compared for a group exposed to the predictions of the AI versus a non-exposed control group receiving standard care [5, 25, 68, 86, 87]. The literature identifies a randomized controlled trial (RCT) as the ideal comparative study design, randomizing patients individually or per cluster [5, 15, 49, 68, 86]. However, this may require more patients and might not always be feasible. Alternative designs are stepped-wedge trials [15, 19, 86], before-after studies [86], and observational studies [5, 19, 57, 68, 86]. For some applications (like imaging technology), a multiple reader multiple case study design is also possible [48], in which the effect of the AIPM on decision making is measured by assessing the differences in discrimination (see phase 2 and 3) with and without the tool. Decision Analytical Modelling may give an initial estimate of clinical utility before commencing a full-blown impact study (see phase 2 and 3) [68, 86].

Trial outcomes can differ across domains and applications. The most mentioned trial outcomes consist of clinical outcomes or patient-reported outcomes [5, 18, 20, 68, 72, 86, 87] followed by cost effectiveness of care [5, 18, 20, 86, 87] and changes in decision making and workflow [5, 20, 68, 86]. Additional trial outcomes are patient experience [20, 57, 87], user satisfaction and engagement [87], and changes in patient (healthy) behavior [87]. It is advised that trial outcomes are also evaluated per clinically relevant user group [12] or per affected non-user group (also in terms of algorithmic bias) [12, 26, 91].

It is recommended that findings are communicated in an understandable and meaningful way to healthcare professionals, but also to administrators and policymakers [56]. AIPM specific guidelines have been developed as extensions to the CONSORT and SPIRIT guidelines for reporting on clinical trials and their protocols respectively [64, 65]. Peer-reviewed open access publication may increase trust and facilitate adoption of the AIPM in a wider clinical community [15].

2.4.6 Phase 6. Implementation and use in daily healthcare practice

Clinical implementation. Clinical implementation consists of all the steps that are necessary to deploy the AIPM in the healthcare environment outside of the clinical trial setting (see phase 5). The literature strongly recommends to state the necessary conditions for deployment before proceeding with the implementation [9, 19, 20, 27, 88]. For example, the AIPM system might require dedicated and locally available hardware [7].

Although not always feasible, the integration of an AIPM directly into the existing medical workflow is preferred [7, 19, 59, 68]. This could for example involve direct integration into the EHR. Moreover, the user is urged to explicitly disclose what part of decision making might be affected by AIPM predictions [26, 27, 43, 62, 63, 88].

To further facilitate the implementation and consecutive monitoring, the literature recommends automatic AIPM deployment (moving software from testing to production environments with automated processes) and the facilitation of shadow deployment [67, 91], which enables prospective local validation (see phase 3) of new versions and updates [19]. Enabling the automatic roll-back for production models is also advised to address real-time operating risks (see phase 4) [67]. Moreover, a procedure to safely abort an operation is highly recommended when the system should stop being used due to a security breach or safety risk [23, 27, 62, 79]. Comparable to the feasibility study of phase 5, pilot studies are recommended to examine the potential pitfalls during implementation, considering both software and hardware issues [8, 16, 72].

Lastly, Institutions and implementers are encouraged to disclose their innovation pathway, including the routes to commercialization [18]. The risks, investments, roles, and responsibilities of the different parties may inform the allocation of benefits in a commercial arrangement [18, 20]. Albeit sparse, [87] provide good guidance on performing economic impact analysis.

Maintenance and updating. Although maintenance is essential to AIPMs (and their software) that are highly dependable on changes in the external world, little guidance can be found on this topic. Developers are recommended to regularly update their AIPMs over time to improve the AIPM's predictive performance as new improvements become available and to mitigate dataset shift [8, 19, 23]. It is advised to pay special attention to the safe and automatic updating of mature systems involving many configurations for many similar models [71]. Note that updating the AIPM may involve recertification. The USA Food and Drug Administration is currently working on a framework that allows for repeated updating of an AIPM without repeated recertification through a change control plan [122].

Education. Education involves the training of end users in the correct use of the AIPM. The literature recommends the general education of end users, often healthcare professionals, on the probabilistic nature [22, 23, 26, 43] and the limitations of AIPMs [22, 43]. This may involve the development of a general AI curriculum for medical students and healthcare professionals.

Application specific training is also advised. The end user may for example be educated on the underlying assumptions of the AIPM [58, 68], its legal framework [27], benefits [20, 27, 58], risks and (technical) limitations [14, 22, 27, 58, 62]. Providing the end user with examples of incorrectly classified cases could help in creating an understanding of the strengths and limitations of the AIPM [13]. Moreover, it is recommended to regularly repeat the training on the correct use of the AIPM [12, 14, 27, 58, 62] and the appropriate response to security breaches [23, 35]. For example, end users may be made aware of the possibility of automation bias and trained to maintain vigilance [22, 27, 56, 88, 91].

When the end user (healthcare professional) and AIPM subject (patient) are different people, as is often the case for AIPMs in healthcare, the literature recommends to train the healthcare professional to explain one's AIPM-supported decisions to their patient [22].

Monitoring and auditing. Monitoring refers to the post-deployment evaluation of the behavior of an AIPM throughout its lifecycle [8, 23, 24, 27, 56, 62, 64, 67, 72, 80, 91]. It is performed by the developer and implementers at the implementation site. Auditing refers to periodic quality control checks of the AIPM (and all of its monitoring aspects) performed by an independent third party [27, 58, 62, 91]. Among other things, It will aid the detection of failures and near misses and through this strengthen the risk management and security of an AIPM [35, 58].

Several aspects of AIPM functioning can be monitored as identified in the literature. These may for example consist of predictive performance and other model outputs [8, 15, 27, 56, 63, 79, 80], distribution of predicted versus observed labels [71], reliability and reproducibility [8, 27, 62], types and severity of errors [56], changes in risk [80], quality of the input data [27, 56, 63, 71, 87], quality of the label [91], case-mix factors [72, 91], accessibility and integration of the model [56], use of the AIPM recommendations [56, 63, 87], user satisfaction and user feedback [8, 15, 56, 79, 87], and (clinical) outcomes [27, 56, 80, 87].

Several monitoring aspects are highlighted in the literature that deserve additional scrutiny. The monitoring of the fairness of an AIPM throughout its lifecycle is often mentioned [12, 15, 20, 23, 26, 27, 63], for example by recording false positive and false negative prediction rates sliced across different subgroups [27, 28, 79, 91]. Second, the monitoring of dataset shift is also repeatedly mentioned in the literature [5, 8, 22, 72, 79, 91]. Dataset shift is a change in the composition of the input data caused by changes in clinical or operational practices over time that can lead to the deterioration of AIPM performance. It can for example be measured by an increase in classification errors over time [23]. It can be mitigated by retraining or updating of the AIPM [72]. One last aspect is the monitoring of feedback loops [27]. They originate when an AIPM is modelled on care delivery features that in turn might be affected by the outcomes of an AIPM.

It is advised to develop integrated mechanisms to facilitate real-time monitoring available at the start of implementation [18, 71]. Implementers are encouraged to clearly define the context and boundaries within which the monitoring is to be performed [56]. Specifying the type of oversight is also recommended, e.g., human-in-the-loop, human-on-the-loop, or human-in-command [27]. Some sources suggest the frequency of the monitoring should be proportional to the AIPM's risks [22, 23, 91]: the higher the risk to the welfare of the patient, the higher the monitoring frequency should be. One source suggests frequent monitoring may be less important for AIPMs solely based on causal mechanisms as they are less likely to change over time [24].

In terms of auditing, the literature recommends the installation of a comprehensive auditability framework [8, 22, 58] and an audit trail [28, 48, 62], in which the AIPM's predictions, model version, input data, and use practices are methodologically logged and made available to interested third parties [22, 27, 33, 35, 58, 61, 62, 67, 91].

Implementers are advised to define mitigation pathways as part of the monitoring and auditing plan to deal with incidents [22, 35, 71, 79]. This may for example involve the regular reporting on failures and near misses and the organization of meetings to discuss incidents [58]. Moreover, the literature states that mitigation could and sometimes should lead to a change in the AIPM's design or use practices, for example an adjustment in the instructions for use, a re-evaluating of the stakeholder impact assessment, or a model update [22, 72, 80].

2.4.7 Current gaps and future perspectives

We identified several important aspects of the AIPM development, evaluation and implementation cycle for which clear guidance was missing in the literature. First, guidance is lacking on the requirements to be fulfilled during the assessment of the medical problem and context. In other words, what aspects of a medical or healthcare problem and setting make the introduction of an AIPM likely to result in better patient care, and when are conditions sufficient to initiate AIPM development? Guidance is also missing on the a priori estimation of a minimum sample size for AIPM development for semi-supervised approaches, and for certain commonly used groups of ML modeling techniques such as decision-trees (e.g., random forests) and deep learning (e.g., convolutional neural networks).

Across all phases, several methodologies and quality criteria were identified to address ethical issues such as algorithmic bias, privacy preservation, and interpretable AI. However, the relevance of these issues for different healthcare domains might differ and so will the preferred definitions, metrics, and techniques to describe and mitigate them. As domain specific guidelines were not the primary focus of this investigation, we cannot with certainty comment on the general absence of such guidelines. Nevertheless, we would advise individual healthcare domains to scrutinize the currently available guidance and, when necessary, address these ethical issues across the AIPM development, evaluation and implementation cycle for their respective settings.

Another aspect for which guidance was limited, is the combination of different data sources (e.g., from different registries and collection sites), and data modalities (e.g. imaging data, electrophysiological data, and lab results) for AIPM development. Although methodological studies exist for various combinations, further research on best practices is needed. Also, current guidance is primarily focused on binary outcomes (e.g., mortality), and guidance is missing on other outcome types (e.g., multinomial, ordinal, hierarchical or sequential outcomes).

Although many standards exist for software security, it is unclear whether they suffice to address cyberattacks particularly geared at AIPMs. Experience with AIPM security in practice and experimentation with the insulation of AIPMs against different types of cyberattack in preclinical settings will help to clarify this. Also, more guidance on the unique aspects of AIPM-specific human-AI interaction design is needed. This will for example entail the presentation of and interaction with probabilistic outcomes and the impact of model interpretability on end users.

Much more guidance is needed addressing how to integrate the AIPM into the current healthcare or clinical workflow. More guidance is also required specifying what design and execution of the feasibility and impact studies are needed, and how to report such studies.

Moreover, guidance is needed regarding the assessment of the cost effectiveness of AIPMs. AIPMs differ from other health technologies and are likely to affect healthcare differently, which should be reflected in their cost effectiveness assessments (as was done for the guidance on impact studies).

We described recommendations regarding the responsibilities of different parties (developers, end-users, organizations) involved with AIPM development and deployment as described in the identified literature (e.g., risk assessment, incident reporting, patient privacy). However, more work is needed addressing the proper distribution of accountability across all involved parties, which may in turn inform institutional governance.

Lastly, guidance is needed on (long-term) maintenance aspects, on dataset shift (and how to mitigate it), and on the frequency and necessity of local validation, recalibration (updating), and retraining. As more and more AIPMs will be implemented into healthcare practice in the coming years, this practical experience can be used to inform these aspects.

2.5 DISCUSSION

This scoping review provides an easy-to-use overview and summary of the currently available actionable guidelines and quality criteria driven by the six phases of the AIPM development, evaluation, and implementation cycle: (1) data preparation, (2) AIPM development, (3) AIPM validation, (4) software development, (5) AIPM impact assessment, and (6) AIPM implementation into daily healthcare practice. Guidance was structured in specific topics and mapped to the different phases and we provided an overview of the current gaps in this guidance.

To appreciate our scoping review and suggested framework of six phases several issues need to be addressed. First, our definitions of ‘actionable’ guidance as an inclusion criterion and the defined six phases are somewhat arbitrary and mainly informed by vast experience with and guidance on developing, evaluating, and implementing prediction models in healthcare. Individual AIPM applications may deviate from the structure presented here. Nevertheless, we believe the phases and their associated topics will translate to most AIPM projects and are in agreement with other phases formulated in the literature [4, 5, 7, 22]. Also, the structure provided by the six phases, and our focus on actionability form two strengths of this scoping review and produce a comprehensible and easy-to-use overview of practical recommendations for those involved in the AIPM development, evaluation and implementation cycle. This sets our review apart from other work that was previously undertaken (e.g., [33, 123, 124]).

Second, the literature databases and sources we used mostly contain scientific literature and only English documents were included in the final search (translations were also considered). This may have biased our results towards academic sources and English-speaking countries of origin. To combat this, we identified additional grey literature through consultation with AI experts and a thorough screening of citations in the included literature. As a result, a substantial number of our included sources can be considered grey literature. Moreover, due to our extensive search, the current summary of available guidelines and quality criteria is comprehensive.

Lastly, the expert group consulted was a convenience sample, resulting in experts predominantly working in the Netherlands. Diversity was obtained by inviting experts with different occupations (e.g., healthcare professionals, data scientists, statisticians, engineers), from different healthcare domains (e.g., radiology, internal medicine, intensive care, primary care, family medicine), and from both academia and industry.

In conclusion, a substantial number of studies provide guidelines and quality criteria pertaining to the AIPM development, evaluation, and implementation cycle, which can be grouped in six well-defined phases. While the opportunities of AIPMs in healthcare are undeniable, the growing interest in these techniques requires careful quality and applicability assessment to guarantee their safety and (cost-)effectiveness before they are used and disseminated in healthcare. This review can serve as the basis for a structured quality assessment framework. Several gaps in the literature were identified where more research is needed. Additional domain and technology specific studies may be necessary and more practical experience with implementing AIPMs is needed to inform further guidance.

2.6 Acknowledgements

We thank Rene Aarnink, Martijn Bauer, Giovanni Cinà, André Dekker, Bart Geerts, Evangelos Kanoulas, Floor van Leeuwen, Wiro Niessen, Daniel Oberski, Egge van der Poel, Maarten de Rijke, Stephan Romeijn, Bart-Jan Verhoeff for

providing and pointing to relevant guidance literature in the field. This research was funded by the Ministry of Health, Welfare and Sport.

REFERENCES

1. van Smeden, M., J.B. Reitsma, R.D. Riley, G.S. Collins, and K.G.M. Moons, *Clinical prediction models: diagnosis versus prognosis.* Journal of Clinical Epidemiology, 2021. **132**: p. 142-145.
2. Moons, K.G., A.P. Kengne, M. Woodward, et al., *Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker.* Heart, 2012. **98**(9): p. 683-90.
3. Collins, G.S., J.B. Reitsma, D.G. Altman, and K.G.M. Moons, *Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement.* Eur Urol, 2015. **67**(6): p. 1142-1151.
4. Steyerberg, E.W., K.G. Moons, D.A. van der Windt, et al., *Prognosis Research Strategy (PROGRESS) 3: prognostic model research.* PLoS Med, 2013. **10**(2): p. e1001381.
5. Moons, K.G., D.G. Altman, Y. Vergouwe, and P. Royston, *Prognosis and prognostic research: application and impact of prognostic models in clinical practice.* Bmj, 2009. **338**: p. b606.
6. Stevens, L.M., B.J. Mortazavi, R.C. Deo, L. Curtis, and D.P. Kao, *Recommendations for Reporting Machine Learning Analyses in Clinical Research.* Circ Cardiovasc Qual Outcomes, 2020. **13**(10): p. e006556.
7. Weikert, T., M. Francone, S. Abbara, et al., *Machine learning in cardiovascular radiology: ESCR position statement on design requirements, quality assessment, current applications, opportunities, and challenges.* European Radiology, 2021. **31**(6): p. 3909-3922.
8. Larson, D.B., H. Harvey, D.L. Rubin, N. Irani, J.R. Tse, and C.P. Langlotz, *Regulatory Frameworks for Development and Evaluation of Artificial Intelligence&#x-2013;Based Diagnostic Imaging Algorithms: Summary and Recommendations.* Journal of the American College of Radiology, 2021. **18**(3): p. 413-424.
9. Smith, M., A. Sattler, G. Hong, and S. Lin, *From Code to Bedside: Implementing Artificial Intelligence Using Quality Improvement Methods.* Journal of General Internal Medicine, 2021. **36**(4): p. 1061-1066.
10. Luo, W., D. Phung, T. Tran, et al., *Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View.* Journal of Medical Internet Research, 2016. **18**(12).
11. Norgeot, B., G. Quer, B.K. Beaulieu-Jones, et al., *Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist.* Nature Medicine, 2020. **26**(9): p. 1320-1324.
12. Machine Intelligence Garage's Ethics Committee. Ethics Framework. Digital Catapult, 2018. Available from: https://www.migarage.ai/wp-content/uploads/2020/11/MIG_Ethics-Report_2020_v5.pdf.
13. Mongan, J., L. Moy, and C.E. Kahn, *Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers.* Radiology: Artificial Intelligence, 2020. **2**(2): p. e200029.
14. Food and Drug Administration, *Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Notification [510(k)] Submissions*. 2012.

15. *Artificial intelligence in health care: The hope, the hype, the promise, the peril*, ed. M. Matheny, et al. 2019, Washington, DC: National Academy of Medicine.
16. Ray, J.M., R.M. Ratwani, C.A. Sinsky, et al., *Six habits of highly successful health information technology: powerful strategies for design and implementation.* Journal of the American Medical Informatics Association, 2019. **26**(10): p. 1109-1114.
17. Steyerberg, E.W. and Y. Vergouwe, *Towards better clinical prediction models: seven steps for development and an ABCD for validation.* Eur Heart J, 2014. **35**(29): p. 1925-31.
18. Vollmer, S., B.A. Mateen, G. Bohner, et al., *Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness.* Bmj-British Medical Journal, 2020. **368**.
19. Wiens, J., S. Saria, M. Sendak, et al., *Do no harm: a roadmap for responsible machine learning for health care.* Nat Med, 2019. **25**(9): p. 1337-1340.
20. UK Department of Health & Social Care. *A guide to good practice for digital and data-driven health technologies*. 2021; Available from: https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology.
21. Bossuyt, P.M., J.B. Reitsma, D.E. Bruns, et al., *Towards Complete and Accurate Reporting of Studies of Diagnostic Accuracy: The STARD Initiative.* Clinical Chemistry, 2003. **49**(1): p. 1-6.
22. Leslie, D. Understanding artificial intelligende ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute., 2019.
23. Information Commissioner's Office. Guidance on the AI auditing framework: Draft guidance for consultation. 2020. Available from: https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf.
24. Cearns, M., T. Hahn, and B.T. Baune, *Recommendations and future directions for supervised machine learning in psychiatry.* Transl Psychiatry, 2019. **9**(1): p. 271.
25. Nykänen, P., J. Brender, J. Talmon, et al., *Guideline for good evaluation practice in health informatics (GEP-HI).* International Journal of Medical Informatics, 2011. **80**(12): p. 815-827.
26. Global Future Council on Human Rights 2016-2018. How to prevent discriminatory outcomes in machine learning. World Economic Forum, 2018. Available from: http://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf.
27. High-Level Expert Group on Artificial Intelligence. The assessment list for trustworthy artificial intelligence (ALTAI) for self-assessment. European Commission, 2020. Available from: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.
28. Turner Lee, N., P. Resnick, and G. Barton. *Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms*. 2019; Available from: https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.
29. *Complete guide to GDPR compliance*. 2020; Available from: https://gdpr.eu/.

30. *Personal Information Protection and Electronic Documents Act,* Canadian Department of Justice, Editor. 2000.

31. *Health Insurance Portability and Accountability Act of 1996,* U.S. Government, Editor. 1996.

32. Rodríguez, N., G. Stipcich, D. Jiménez, et al., *Federated Learning and Differential Privacy: Software tools analysis, the Sherpa.ai FL framework and methodological guidelines for preserving data privacy.* Information Fusion, 2020. **64**.

33. Ryan, M. and B.C. Stahl, *Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications.* Journal of Information, Communication and Ethics in Society, 2021. **19**(1): p. 61-86.

34. Liaw, S.T., H. Liyanage, C. Kuziemsky, et al., *Ethical Use of Electronic Health Record Data and Artificial Intelligence: Recommendations of the Primary Care Informatics Working Group of the International Medical Informatics Association.* Yearb Med Inform, 2020. **29**(1): p. 51-57.

35. Datatilsynet. Software development with Data Protection by Design and by Default. The Norwegian Data Protection Authority, 2017. Available from: https://www.datatilsynet.no/en/about-privacy/virksomhetenes-plikter/innebygd-personvern/data-protection-by-design-and-by-default/?print=true.

36. Sartor, G. and F. Lagioia. The impact of the General Data Protection Regulation (GDPR) on artificial intelligence. Panel for the Future of Science and Technology, 2020. Available from: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf.

37. Gebru, T., J. Morgenstern, B. Vecchione, et al., *Datasheets for Datasets*. 2020.

38. Forcier, M.B., H. Gallois, S. Mullan, and Y. Joly, *Integrating artificial intelligence into health care through data access: can the GDPR act as a beacon for policymakers?* Journal of Law and the Biosciences, 2019. **6**(1): p. 317-335.

39. *The Privacy Act,* O.o.t.P.C.o. Canada, Editor. 1985.

40. Datatilsynet. Artificial intelligence and privacy. The Norwegian Data Protection Authority, 2018. Available from: https://www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf.

41. Information Commissioner's Office. *ICO and the Turing consultation on explaining AI decisions guidance*. 2020; Available from: https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-and-the-turing-consultation-on-explaining-ai-decisions-guidance/.

42. Arrieta, A.B., N. Diaz-Rodriguez, J. Del Ser, et al., *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.* Information Fusion, 2020. **58**: p. 82-115.

43. Benjamins, R., A. Barbado, and D. Sierra. *Responsible AI by Design in Practice*. in *AAAI Fall Symposium*. 2019.

44. Information Commissioner's Office. Anonymisation: Managing data protection risk code of practice. 2012. Available from: https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf.

45. *California Consumer Privacy Act of 2018*, S.o. California, Editor. 2018.

46. Bhaskar, H., D.C. Hoyle, and S. Singh, *Machine learning in bioinformatics: a brief survey and recommendations for practitioners.* Comput Biol Med, 2006. **36**(10): p. 1104-25.

47. Riley, R.D., J. Ensor, K.I.E. Snell, et al., *Calculating the sample size required for developing a clinical prediction model.* BMJ, 2020. **368**: p. m441.
48. Food and Drug Administration, *Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data in Premarket Notification (510(k)) Submissions*. 2020.
49. Steyerberg, E.W., *Clinical Prediction Models*, ed. M. Gail, M.S. Jonathan, and B. Singer. 2009, Cham, Switzerland: Springer Nature.
50. Pineau, J., P. Vincent-Lamarre, K. Sinha, et al., *The Machine Learning Reproducibility Checklist*, in *Neural Information Processing Systems*. 2020.
51. Christodoulou, E., M. van Smeden, M. Edlinger, et al., *Adaptive sample size determination for the development of clinical prediction models.* Diagnostic and Prognostic Research, 2021. **5**(1): p. 6.
52. Mukherjee, S., P. Tamayo, S. Rogers, et al., *Estimating Dataset Size Requirements for Classifying DNA Microarray Data.* Journal of Computational Biology, 2003. **10**(2): p. 119-142.
53. Vergouwe, Y., E.W. Steyerberg, M.J.C. Eijkemans, and J.D.F. Habbema, *Substantial effective sample sizes were required for external validation studies of predictive logistic regression models.* Journal of Clinical Epidemiology, 2005. **58**(5): p. 475-483.
54. Riley, R.D., T.P.A. Debray, G.S. Collins, et al., *Minimum sample size for external validation of a clinical prediction model with a binary outcome.* Statistics in Medicine, 2021. **40**(19): p. 4230-4251.
55. Archer, L., K.I.E. Snell, J. Ensor, M.T. Hudda, G.S. Collins, and R.D. Riley, *Minimum sample size for external validation of a clinical prediction model with a continuous outcome.* Statistics in Medicine, 2021. **40**(1): p. 133-146.
56. Magrabi, F., E. Ammenwerth, J.B. McNair, et al., *Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications.* Yearb Med Inform, 2019. **28**(1): p. 128-134.
57. Park, S.H. and K. Han, *Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction.* Radiology, 2018. **286**(3): p. 800-809.
58. Shneiderman, B., *Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems.* ACM Trans. Interact. Intell. Syst., 2020. **10**(4): p. Article 26.
59. Wang, F. and A. Preininger, *AI in Health: State of the Art, Challenges, and Future Directions.* Yearb Med Inform, 2019. **28**(1): p. 16-26.
60. Van Calster, B., D.J. McLernon, M. van Smeden, et al., *Calibration: the Achilles heel of predictive analytics.* BMC Medicine, 2019. **17**(1): p. 230.
61. Diakopoulos, N., S. Friedler, M. Arenas, et al. *Principles for Accountable Algorithms and a Social Impact Statement for Algorithms*. Available from: https://www.fatml.org/resources/principles-for-accountable-algorithms.
62. High-Level Expert Group on Artificial Intelligence. Draft ethics guidelines for trustworthy AI. European Commission, 2018. Available from: https://www.euractiv.com/wp-content/uploads/sites/2/2018/12/AIHLEGDraftAIEthicsGuidelinespdf.pdf.

63. Monetary Authority of Singapore. Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector. 2019. Available from: https://www.mas.gov.sg/~/media/MAS/News%20and%20Publications/Monographs%20and%20Information%20Papers/FEAT%20Principles%20Final.pdf.

64. Liu, X., S.C. Rivera, D. Moher, M.J. Calvert, and A.K. Denniston, *Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension.* Bmj, 2020. **370**: p. m3164.

65. Rivera, S.C., X.X. Liu, A.W. Chan, et al., *Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension.* Nature Medicine, 2020. **26**(9): p. 1351-1363.

66. Alonso, O., *Challenges with Label Quality for Supervised Learning.* Acm Journal of Data and Information Quality, 2015. **6**(1).

67. Serban, A., K.v.d. Blom, H. Hoos, and J. Visser, *Adoption and Effects of Software Engineering Best Practices in Machine Learning*, in *Proceedings of the 14th ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*. 2020, Association for Computing Machinery: Bari, Italy. p. Article 3.

68. Kappen, T.H., W.A. van Klei, L. van Wolfswinkel, C.J. Kalkman, Y. Vergouwe, and K.G.M. Moons, *Evaluating the impact of prediction models: lessons learned, challenges, and recommendations.* Diagn Progn Res, 2018. **2**: p. 11.

69. Poldrack, R.A., G. Huckins, and G. Varoquaux, *Establishment of Best Practices for Evidence for Prediction: A Review.* JAMA Psychiatry, 2020. **77**(5): p. 534-540.

70. Kaufman, S., S. Rosset, C. Perlich, and O. Stitelman, *Leakage in data mining: Formulation, detection, and avoidance.* ACM Trans. Knowl. Discov. Data, 2012. **6**(4): p. Article 15.

71. Sculley, D., G. Holt, D. Golovin, et al., *Hidden technical debt in Machine learning systems*, in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. 2015, MIT Press: Montreal, Canada. p. 2503–2511.

72. Kelly, C.J., A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, *Key challenges for delivering clinical impact with artificial intelligence.* BMC Med, 2019. **17**(1): p. 195.

73. Miller, T., *Explanation in artificial intelligence: Insights from the social sciences.* Artificial Intelligence, 2019. **267**: p. 1-38.

74. Huang, S.C., A. Pareek, S. Seyyedi, I. Banerjee, and M.P. Lungren, *Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines.* NPJ Digit Med, 2020. **3**: p. 136.

75. Molnar, C., *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2019.

76. Aliferis, C.F., A. Statnikov, and I. Tsamardinos, *Challenges in the analysis of mass-throughput data: a technical commentary from the statistical machine learning perspective.* Cancer Inform, 2007. **2**: p. 133-62.

77. Eggensperger, K., M. Lindauer, and F. Hutter, *Pitfalls and best practices in algorithm configuration.* J. Artif. Int. Res., 2019. **64**(1): p. 861–893.

78. Altman, D.G., Y. Vergouwe, P. Royston, and K.G.M. Moons, *Prognosis and prognostic research: validating a prognostic model.* BMJ, 2009. **338**: p. b605.

79. Google AI. *Responsible AI practices*. 2021; Available from: https://ai.google/responsibilities/responsible-ai-practices/.
80. Food and Drug Administration, *Software as a Medical Device (SAMD): Clinical Evaluation*. 2017.
81. Wynants, L., M. van Smeden, D.J. McLernon, et al., *Three myths about risk thresholds for prediction models.* BMC Medicine, 2019. **17**(1): p. 192.
82. Kaur, H., H.S. Pannu, and A.K. Malhi, *A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions.* Acm Computing Surveys, 2019. **52**(4).
83. Van Calster, B., M. van Smeden, B. De Cock, and E.W. Steyerberg, *Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study.* Statistical Methods in Medical Research, 2020. **29**(11): p. 3166-3178.
84. Riley, R.D., K.I.E. Snell, G.P. Martin, et al., *Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small.* Journal of Clinical Epidemiology, 2021. **132**: p. 88-96.
85. Pozzolo, A.D., O. Caelen, R.A. Johnson, and G. Bontempi. *Calibrating Probability with Undersampling for Unbalanced Classification*. in *2015 IEEE Symposium Series on Computational Intelligence*. 2015.
86. Moons, K.G., A.P. Kengne, D.E. Grobbee, et al., *Risk prediction models: II. External validation, model updating, and impact assessment.* Heart, 2012. **98**(9): p. 691-8.
87. National Institute for Health and Care Excellence. Evidence standards framework for digital health technologies. 2018. Available from: https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies.
88. Berscheid, J. and F. Roewer-Despres, *Beyond transparency: a proposed framework for accountability in decision-making AI systems.* AI Matters, 2019. **5**(2): p. 13–22.
89. Su, T.L., T. Jaki, G.L. Hickey, I. Buchan, and M. Sperrin, *A review of statistical updating methods for clinical prediction models.* Stat Methods Med Res, 2018. **27**(1): p. 185-197.
90. Jenkins, D.A., G.P. Martin, M. Sperrin, et al., *Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems?* Diagnostic and Prognostic Research, 2021. **5**(1): p. 1.
91. McCradden, M.D., S. Joshi, J.A. Anderson, M. Mazwi, A. Goldenberg, and R. Zlotnik Shaul, *Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning.* J Am Med Inform Assoc, 2020. **27**(12): p. 2024-2027.
92. International Organization for Standardization. *Artificial intelligence (ISO/IEC JTC 1/SC 42)*. 2017; Available from: https://www.iso.org/committee/6794475.html.
93. Institute of Electrical and Electronics Engineers. *IEEE Approved Draft Model Process for Addressing Ethical Concerns During System Design (IEEE 7000-2021)*. 2021; Available from: https://standards.ieee.org/standard/7000-2021.html.
94. HL7. *FHIR*. 2019; Available from: http://hl7.org/fhir/.
95. HL7. 2021; Available from: http://www.hl7.org/.

96. International Organization for Standardization. *Health informatics - Personal health device communication - Part 10418: Device specialization - International Normalized Ratio (INR) monitor (ISO/IEEE 11073-10418:2014)*. 2014; Available from: https://www.iso.org/standard/61897.html.
97. International Organization for Standardization. *Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems (ISO 9241-210:2019)*. 2019; Available from: https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-2:v1:en.
98. International Organization for Standardization. *Medical devices - Part 1: Application of usability engineering to medical devices (IEC 62366-1:2015)*. 2015; Available from: https://www.iso.org/standard/63179.html.
99. Amershi, S., D. Weld, M. Vorvoreanu, et al., *Guidelines for Human-AI Interaction*, in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, Association for Computing Machinery: Glasgow, Scotland Uk. p. Paper 3.
100. eHealth Observatory. Canada Health Infoway Benefits Evaluation Indicators. 2012.
101. Sendak, M.P., M. Gao, N. Brajer, and S. Balu, *Presenting machine learning model information to clinical end users with model facts labels.* npj Digital Medicine, 2020. **3**(1): p. 41.
102. Medicines & Healthcare products Regulatory Agency, *Guidance on applying human factors and usability engineering to medical devices including drug-device combination products in Great Britain*. 2021.
103. Food and Drug Administration, *Applying human factors and usability engineering to medical devices: Guidance for industry and food and drug administrations taff*. 2016.
104. Council of the European Union, *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 (Medical Device Regulation)*. 2017.
105. Council of the European Union, *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.
106. National Electrical Manufacturers Association. *Manufacturer disclosure statement for medical device security (ANSI/NEMA NH 1-2019)*. 2019; Available from: https://www.nema.org/standards/view/manufacturer-disclosure-statement-for-medical-device-security.
107. Royal Netherlands Standardization Institute. *Health informatics - Information security management in healthcare - Part 1: Management system (NEN 7510)*. 2020; Available from: https://www.nen.nl/en/nen-7510-1-2017-a1-2020-nl-267179.
108. European Commission, *Guidance on Cybersecurity for medical devices*. 2020.
109. UL Standards. *ANSI/CAN/UL Standard for software cybersecurity for network-connectable products, part1: General requirements (ANSI/CAN/UL standard 2900-1)*. 2017; Available from: https://standardscatalog.ul.com/ProductDetail.aspx?productId=UL2900-1.
110. International Medical Device Regulators Forum, *Principles and practices for medical device cypersecurity*. 2020.

111. Food and Drug Administration, *Response to NIST workshop and call for position papers on standards and guidelines to enhance software supply chain security*. 2021.

112. International Organization for Standardization. *Information technology, cybersecurity and privacy protection - Cybersecurity framework development guidelines (ISO/IEC TS 27110:2021)*. 2021; Available from: https://www.iso.org/standard/72435.html.

113. International Organization for Standardization. *Information technology - Security techniques - Guidelines for cybersecurity (ISO/IEC 27032:2012)*. 2012; Available from: https://www.iso.org/standard/44375.html.

114. International Organization for Standardization. *Information technology - Security techniques - Governance of information security (ISO/IEC 27014:2013)*. 2013; Available from: https://www.iso.org/standard/43754.html.

115. International Organization for Standardization. *Information technology - Security techniques - Code of practice for information security controls (ISO/IEC 27002:2013)*. 2013; Available from: https://www.iso.org/standard/54533.html.

116. Food and Drug Administration, *Postmarket surveillance under section 522 of the federal food, drug, and cosmetic act*. 2016.

117. International Organization for Standardization. *Medical device software - Software life cycle processes (IEC 62304:2006)*. 2006; Available from: https://www.iso.org/obp/ui/#iso:std:iec:62304:ed-1:v1:en.

118. International Organization for Standardization. *Health software - Part 1: General requirements for product safety (IEC 82304-1:2016)*. 2016; Available from: https://www.iso.org/standard/59543.html.

119. International Organization for Standardization. *Medical devices - Application of risk management to medical devices (ISO 14971:2019)*. 2019; Available from: https://www.iso.org/standard/72704.html.

120. Food and Drug Administration, *General principles of software validation*. 2002.

121. Food and Drug Administration, *Off-the-shelf software use in medical devices*. 2019.

122. Food and Drug Administration, *Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD)*. 2019.

123. Hagendorff, T., *The Ethics of AI Ethics: An Evaluation of Guidelines.* Minds and Machines, 2020. **30**(1): p. 99-120.

124. Jobin, A., M. Ienca, and E. Vayena, *The global landscape of AI ethics guidelines.* Nature Machine Intelligence, 2019. **1**(9): p. 389-399.

SUPPLEMENTARY MATERIAL

Table S1 The search query

Database	Search string
PubMed	(("data cleaning" OR "data preparation" OR "preprocessing" OR "pre-processing" OR "design" OR "missing data" OR "outlier detection" OR "data harmonization" OR "de-identification" OR "anonymization" OR "predictor selection " OR "feature selection" OR "feature extraction" OR "selection bias" OR "annotation" OR "sample size" OR "data privacy" OR "model development" OR "model architectures" OR "explainability" OR "hyper-parameter" OR "hyperparameter" OR "model training" OR "model fitting" OR "optimization" OR "model updating" OR "parameter-sharing" OR "distant supervision" OR "weak supervision" OR "interpretability" OR "interpretable" OR "evaluation" OR "calibration" OR "discrimination" OR "evaluation metric" OR "evaluation measure" OR "model accuracy" OR "risk of bias" OR "prediction performance" OR "generalization error" OR "prediction error" OR "validation" OR "net-benefit" OR "precision" OR "software" OR "software quality assurance" OR "SQA" OR "software quality control" OR "SQC" OR "software as a medical device" OR "SaMD" OR "software compliance" OR "anti-patterns" OR "design patterns" OR "software architecture" OR "front-end design" OR "presentation layer" OR "software testing" OR "software security" OR "Software as a service" OR "SaaS" OR "miminum viable product" OR "impact assessment" OR "outcome assessment" OR "validation" OR "clinical performance" OR "clinical investigation" OR "external validation" OR "RCT" OR "randomized controlled trial" OR "randomized clinical trial" OR "random control trial" OR "technology assessment" OR "HTA" OR "clinical impact" OR "pilot study" OR "clinical benefit" OR "clinical evaluation" OR "cost-effectiveness" OR "generalizability" OR "explainability" OR "clinical benchmarking" OR "study design" OR "fairness" OR "bias" OR "qualitative evaluation" OR "implementation" OR "scalability" OR "integration" OR "calibration" OR "transfer learning" OR "usability" OR "patient satisfaction" OR "satisfaction" OR "interoperability" OR "user friendly" OR "ethics" OR "ethical" OR "jurisprudence" OR "legislation" OR "legal" OR "law" OR "diffusion" OR "application" OR "dissemination" OR "real-world performance" OR "real world performance" OR "monitoring" OR "clinical practice" OR "education") AND ("artificial intelligence"[Title] OR "machine intelligence"[Title] OR "machine learning"[Title] OR "deep learning"[Title] OR "prediction model"[Title] OR "neural network"[Title] OR "support vector machines"[Title] OR "natural language processing"[Title] OR "computer vision"[Title] OR "supervised learning"[Title] OR "unsupervised learning"[Title] OR "reinforcement learning"[Title] OR "statistical learning"[Title] OR "computational intelligence"[Title] OR "computer reasoning"[Title] OR "AI"[Title] OR "computer heuristics"[Title] OR "expert systems"[Title])) AND ("recommendations"[Title] OR "challenges"[Title] OR "guideline"[Title] OR "a guide"[Title] OR "guidelines"[Title] OR "practice guideline"[Title] OR "practice guidelines"[Title] OR "quality norm"[Title] OR "quality of care"[Title] OR "quality criteria"[Title] OR "quality instrument"[Title] OR "quality of health care"[Title] OR "healthcare quality"[Title] OR "quality improvement"[Title] OR "quality indicator"[Title] OR "quality indicators"[Title] OR "total quality management"[Title] OR "best practice"[Title] OR "code of conduct"[Title] OR "reporting standard"[Title] OR "good machine learning practice"[Title] OR "best practices"[Title] OR "framework"[Title] OR "guidance"[Title] OR "strategies for"[Title] OR "statement"[Title])

Table S1 The search query (continued)

Database	Search string
Web of Science	TS=(("guideline" OR "a guide" OR "guidelines" OR "practice guideline" OR "practice guidelines" OR "quality norm" OR "quality of care" OR "quality criteria" OR "quality instrument" OR "quality of health care" OR "healthcare quality" OR "quality improvement" OR "quality indicator" OR "quality indicators" OR "total quality management" OR "best practice" OR "code of conduct" OR "reporting standard" OR "good machine learning practice" OR "best practices" OR "framework" OR "guidance" OR "strategies for" OR "statement" OR "recommendations" OR "challenges") AND ("artificial intelligence" OR "machine intelligence" OR "machine learning" OR "deep learning" OR "prediction model" OR "neural network" OR "support vector machines" OR "natural language processing" OR "computer vision" OR "supervised learning" OR "unsupervised learning" OR "reinforcement learning" OR "statistical learning" OR "computational intelligence" OR "computer reasoning" OR "AI" OR "computer heuristics" OR "expert systems")) AND AB=(("data cleaning" OR "data preparation" OR "preprocessing" OR "pre-processing" OR "design" OR "missing data" OR "outlier detection" OR "data harmonization" OR "de-identification" OR "anonymization" OR "predictor selection " OR "feature selection" OR "feature extraction" OR "selection bias" OR "annotation" OR "sample size" OR "data privacy" OR "model development" OR "model architectures" OR "explainability" OR "hyper-parameter" OR "hyperparameter" OR "model training" OR "model fitting" OR "optimization" OR "model updating" OR "parameter-sharing" OR "distant supervision" OR "weak supervision" OR "interpretability" OR "interpretable" OR "evaluation" OR "calibration" OR "discrimination" OR "evaluation metric" OR "evaluation measure" OR "model accuracy" OR "risk of bias" OR "prediction performance" OR "generalization error" OR "prediction error" OR "validation" OR "net-benefit" OR "precision" OR "software" OR "software quality assurance" OR "SQA" OR "software quality control" OR "SQC" OR "software as a medical device" OR "SaMD" OR "software compliance" OR "anti-patterns" OR "design patterns" OR "software architecture" OR "front-end design" OR "presentation layer" OR "software testing" OR "software security" OR "Software as a service" OR "SaaS" OR "minimum viable product" OR "impact assessment" OR "outcome assessment" OR "validation" OR "clinical performance" OR "clinical investigation" OR "external validation" OR "RCT" OR "randomized controlled trial" OR "randomized clinical trial" OR "random control trial" OR "technology assessment" OR "HTA" OR "clinical impact" OR "pilot study" OR "clinical benefit" OR "clinical evaluation" OR "cost-effectiveness" OR "generalizability" OR "explainability" OR "clinical benchmarking" OR "study design" OR "fairness" OR "bias" OR "qualitative evaluation" OR "implementation" OR "scalability" OR "integration" OR "calibration" OR "transfer learning" OR "usability" OR "patient satisfaction" OR "satisfaction" OR "interoperability" OR "user friendly" OR "ethics" OR "ethical" OR "jurisprudence" OR "legislation" OR "legal" OR "law" OR "diffusion" OR "application" OR "dissemination" OR "real-world performance" OR "real world performance" OR "monitoring" OR "clinical practice" OR "education") AND (Healthcare domain))

Table S1 The search query (continued)

Database	Search string
ACM Digital Library	[[Publication Title: "recommendations"] OR [Publication Title: "challenges"] OR [Publication Title: "guideline"] OR [Publication Title: "a guide"] OR [Publication Title: "guidelines"] OR [Publication Title: "practice guideline"] OR [Publication Title: "practice guidelines"] OR [Publication Title: "quality norm"] OR [Publication Title: "quality of care"] OR [Publication Title: "quality criteria"] OR [Publication Title: "quality instrument"] OR [Publication Title: "quality of health care"] OR [Publication Title: "healthcare quality"] OR [Publication Title: "quality improvement"] OR [Publication Title: "quality indicator"] OR [Publication Title: "quality indicators"] OR [Publication Title: "total quality management"] OR [Publication Title: "best practice"] OR [Publication Title: "code of conduct"] OR [Publication Title: "reporting standard"] OR [Publication Title: "good machine learning practice"] OR [Publication Title: "best practices"] OR [Publication Title: "framework"] OR [Publication Title: "guidance"] OR [Publication Title: "strategies for"] OR [Publication Title: "statement"]] AND [[Publication Title: "artificial intelligence"] OR [Publication Title: "machine intelligence"] OR [Publication Title: "machine learning"] OR [Publication Title: "deep learning"] OR [Publication Title: "prediction model"] OR [Publication Title: "neural network"] OR [Publication Title: "support vector machines"] OR [Publication Title: "natural language processing"] OR [Publication Title: "computer vision"] OR [Publication Title: "supervised learning"] OR [Publication Title: "unsupervised learning"] OR [Publication Title: "reinforcement learning"] OR [Publication Title: "statistical learning"] OR [Publication Title: "computational intelligence"] OR [Publication Title: "computer reasoning"] OR [Publication Title: "ai"] OR [Publication Title: "computer heuristics"] OR [Publication Title: "expert systems"]] AND [[Full Text: "data cleaning"] OR [Full Text: "data preparation"] OR [Full Text: "preprocessing"] OR [Full Text: "pre-processing"] OR [Full Text: "design"] OR [Full Text: "missing data"] OR [Full Text: "outlier detection"] OR [Full Text: "data harmonization"] OR [Full Text: "de-identification"] OR [Full Text: "anonymization"] OR [Full Text: "predictor selection "] OR [Full Text: "feature selection"] OR [Full Text: "feature extraction"] OR [Full Text: "selection bias"] OR [Full Text: "annotation"] OR [Full Text: "sample size"] OR [Full Text: "data privacy"] OR [Full Text: "model development"] OR [Full Text: "model architectures"] OR [Full Text: "explainability"] OR [Full Text: "hyper-parameter"] OR [Full Text: "hyperparameter"] OR [Full Text: "model training"] OR [Full Text: "model fitting"] OR [Full Text: "optimization"] OR [Full Text: "model updating"] OR [Full Text: "parameter-sharing"] OR [Full Text: "distant supervision"] OR [Full Text: "weak supervision"] OR [Full Text: "interpretability"] OR [Full Text: "interpretable"] OR [Full Text: "evaluation"] OR [Full Text: "calibration"] OR [Full Text: "discrimination"] OR [Full Text: "evaluation metric"] OR [Full Text: "evaluation measure"] OR [Full Text: "model accuracy"] OR [Full Text: "risk of bias"] OR [Full Text: "prediction performance"] OR [Full Text: "generalization error"] OR [Full Text: "prediction error"] OR [Full Text: "validation"] OR [Full Text: "net-benefit"] OR [Full Text: "precision"] OR [Full Text: "software"] OR [Full Text: "software quality assurance"] OR [Full Text: "sqa"] OR [Full Text: "software quality control"] OR [Full Text: "sqc"] OR [Full Text: "software as a medical device"] OR [Full Text: "samd"] OR [Full Text: "software compliance"] OR [Full Text: "anti-patterns"] OR [Full Text: "design patterns"] OR [Full Text: "software architecture"] OR [Full Text: "front-end design"] OR [Full Text: "presentation layer"] OR [Full Text: "software testing"] OR [Full Text: “software security”] OR [Full Text: “software

Table S1 The search query (continued)

Database	Search string
	as a service”] OR [Full Text: “saas”] OR [Full Text: “miminum viable product”] OR [Full Text: “impact assessment”] OR [Full Text: “outcome assessment”] OR [Full Text: “validation”] OR [Full Text: “clinical performance”] OR [Full Text: “clinical investigation”] OR [Full Text: “external validation”] OR [Full Text: “rct”] OR [Full Text: “randomized controlled trial”] OR [Full Text: “randomized clinical trial”] OR [Full Text: “random control trial”] OR [Full Text: “technology assessment”] OR [Full Text: “hta”] OR [Full Text: “clinical impact”] OR [Full Text: “pilot study”] OR [Full Text: “clinical benefit”] OR [Full Text: “clinical evaluation”] OR [Full Text: “cost-effectiveness”] OR [Full Text: “generalizability”] OR [Full Text: “explainability”] OR [Full Text: “clinical benchmarking”] OR [Full Text: “study design”] OR [Full Text: “fairness”] OR [Full Text: “bias”] OR [Full Text: “qualitative evaluation”] OR [Full Text: “implementation”] OR [Full Text: “scalability”] OR [Full Text: “integration”] OR [Full Text: “calibration”] OR [Full Text: “transfer learning”] OR [Full Text: “usability”] OR [Full Text: “patient satisfaction”] OR [Full Text: “satisfaction”] OR [Full Text: “interoperability”] OR [Full Text: “user friendly”] OR [Full Text: “ethics”] OR [Full Text: “ethical”] OR [Full Text: “jurisprudence”] OR [Full Text: “legislation”] OR [Full Text: “legal”] OR [Full Text: “law”] OR [Full Text: “diffusion”] OR [Full Text: “application”] OR [Full Text: “dissemination”] OR [Full Text: “real-world performance”] OR [Full Text: “real world performance”] OR [Full Text: “monitoring”] OR [Full Text: “clinical practice”] OR [Full Text: “education”]] AND [Publication Date: (01/01/2000 TO *)]

Table S2 The consulted experts

Name	Affiliation	Expertise
Maarten de Rijke	University of Amsterdam	Artificial intelligence
Evangelos Kanoulas	University of Amsterdam	Machine learning and statistics
Floor van Leeuwen	Quantib	Medical device regulation
Daniel Oberski	Utrecht University	Machine learning and statistics
Wiro Niessen	Erasmus MC, University Medical Center Rotterdam & Delft University of Technology	Medical image processing
Giovanni Cina	Pacmed	Artificial intelligence
Rene Aarnink	Philips	Artificial intelligence
Anonymous	-	Medical device regulation
Bart-Jan Verhoeff	Expertisecentrum Zorgalgoritmen	Clinical software development
Bart Geerts	Healthplus.ai & Spaarne Gasthuis	Clinical AI implementation
Egge van der Poel	Erasmus Medical Center	Personalized healthcare
Stephan Romeijn	Leiden University Medical Center	Clinical AI implementation
Martijn Bauer	Leiden University Medical Center	Clinical AI implementation
André Dekker	Maastricht University	Clinical data science

Table S3 The mapping from the initially used more fine-grained topics and terms identified in the literature, to more coarse-grained topics used in the review's outline, and their distribution over the different phases from Box 1.

Keywords	Topic	Phase
Problem definition; analysis of the status quo; background study; specification of the clinical setting; identification of stakeholders; the prediction task; motivation; clinical context; clinical workflow; clinical baseline; clinical setting; clinical question	Medical problem and context	Phase 1
GDPR; privacy; informed consent; data governance; de-identification; lawful basis; data minimization; compliance; ethics	Patient privacy	
Sample size	Sample size	
Representative data; study population; spectrum bias	Representativeness	
Missing data; measurement error; data quality; outliers; inter-annotator agreement; labelling; annotation; ground truth quality; data cleaning; data collection; reference standard; outcome measures; reference test	Data quality	
Data cleaning; data preparation; outlier detection; imputation; data wrangling; data fusion; integrating data; fusion; coding predictors	Data preprocessing	
Data standards; interoperability	Data coding standards	
Interpretability; model selection; deep learning; federated learning; model specification; statistical models	Model selection and interpretability	Phase 2
Parameter tuning; hyperparameters; nested cross-validation; development; training; train-test split	Training the AIPM	
Overfitting; dimensionality reduction; feature selection; class imbalance; clustering; regularization; predictor selection	Measures to reduce risk of overfitting	
Fairness; discriminatory bias; equality; algorithmic bias	Measures to identify and prevent discriminatory bias	
Validation; evaluation; metrics; cross-validation; train-test split; evaluation metrics; calibration; discrimination; internal validation	Internal validation	
Reporting; code sharing	Transparency of the modelling process	

Table S3 The mapping from the initially used more fine-grained topics and terms identified in the literature, to more coarse-grained topics used in the review's outline, and their distribution over the different phases from Box 1. (continued)

Keywords	Topic	Phase
Validation; evaluation; metrics; comparing models; evaluation metrics; external validation; prospective vs. retrospective; calibration; discrimination; benchmarking; temporal validation	Validation of the AIPM	Phase 3
Generalizability	Generalizability	
Interoperability; open source; data standards; software design	Interoperability	Phase 4
Human-AI interaction; usability; human-machine interaction; design; user interface; HCI	Human-AI interaction	
Logging; software updating; facilitating monitoring	Facilitating software updating and monitoring	
Security; adversarial attack; model inversion attack	Security	
Security; risks; software design; coding; unit testing; software testing	Software testing	
Feasibility study; clinical utility; patient treatment strategy	Feasibility study	Phase 5
Effectiveness; clinical outcome; impact assessment; clinical impact assessment; RCT; economic impact assessment	Impact study	
Risks; risk management	Risk management	
Implementation; patient-physician relation; integration into clinical workflow; clinical implementation; trust	Clinical implementation	Phase 6
Hardware; maintenance	Maintenance	
Education	Education	
Auditing; data drift; monitoring; concept drift; performance over time; third-party evaluation;	Monitoring and auditing	
Fairness; algorithmic bias; bias; ethics	Algorithmic bias & fairness	Overarching
Reporting; open source; code sharing; transparency; trust	Transparency & openness	
Explainability; saliency mapping; trust	Interpretability	
Accountability; stakeholders; multidisciplinary team	Development team, end users, and stakeholders	
Security; safety; model inversion attack; adversarial attack; trust	Security	
Risks; safety; risk management	Risks	

Table S4 Index on where phase-overarching topics are discussed in the article's summary of the found guidance

	Topic	**Algorithmic bias and fairness**	**Transparency and openness**	**Interpretability**	**Team members, end users, and stakeholders**	**Security**	**Risks**
Phase 1	Medical problem and context	√	√	.	√	.	√
	Patient privacy	.	√	.	√	√	.
	Sample size	.	√	.	.	.	.
	Representativeness	√	√	.	.	.	.
	Data quality	√	√	.	√	.	√
	Data preprocessing	.	√	.	√	.	√
	Data coding standards	.	.	.	.	.	.
Phase 2	Model selection and interpretability	√	√	√	.	√	√
	Training the AIPM	.	√	.	.	.	.
	Internal validation	.	√	.	.	.	√
	Measures to reduce risk of overfitting	.	√	.	√	.	√
	Measures to identify and prevent algorithmic bias	√	√	√	√	.	.
	Transparency of the modelling process	.	√	.	.	.	.
Phase 3	Validation of the AIPM	.	.	.	√	.	√
	Generalizability	√	√	.	.	.	√
Phase 4	Interoperability	.	√	.	.	.	√
	Human-AI interaction	√	√	√	√	.	√
	Facilitating software updating and monitoring	.	√	.	.	.	√
	Security	.	√	.	√	√	√
	Software testing	.	.	.	.	√	√
Phase 5	Feasibility study	.	√	.	√	.	.
	Risk management	.	√	.	.	.	√
	Impact study	√	√	.	.	.	.
Phase 6	Clinical implementation	.	√	.	.	√	√
	Maintenance and updating	.	.	.	.	.	√
	Education	.	√	.	√	√	√
	Monitoring and auditing	√	√	.	√	√	√