August 2020

Guidance and Quality Criteria on AI Prediction Algorithms in Medical Devices: Key Publications

Phase 1-3

Tuur Leeuwenberg[1], Steven Nijman[1], Rene Spijker[1], Thomas Debray[1], Ewoud Schuit[1], Maarten van Smeden[1], Johannes Reitsma[1], Lotty Hooft[1], Carl Moons[1]

Phase 4-6

Anne A. H. de Hond[2,3,4], Hendrikus J. A. van Os[5], Jiska J. Aardoom[5,6], Niels H. Chavannes[5,6]

[1]*Julius Center for Health Sciences and Primary Care, University Medical Center, Utrecht University, Universiteitsweg 100, 3584 CG Utrecht, Netherlands*
[2]*Waardegedreven Zorg & AI, Informatie Technologie & Digitale Innovatie, LUMC*
[3]*Clinical AI Implementation and Research Lab, LUMC*
[4]*Department of Biomedical Data Sciences, LUMC*
[5]*National eHealth Living Lab, LUMC*
[6]*Department of Public Health and Primary Care, LUMC*

Glossary

AI prediction algorithm (AIPA) or models

A data-driven algorithm that provides individual participant- or patient-level predictions of having (diagnosis) or developing (prognosis) a certain outcome (e.g., a certain health outcome), given certain input features (e.g., certain patient characteristics, test results, genetic markers, medical images, or other type of predictors). Such prediction algorithms can include supervised and unsupervised algorithms.

Training data (or development data)

A (participant or patient) dataset used to develop an AIPA.

Test data (or validation data)

A (participant or patient) dataset used to quantify and validate the AIPA's predictive performance (has no overlap with the training data).

Tabular data

Data that can be structured naturally into rows and columns.

Unstructured data

Data which has no pre-defined data model that describes all possible data values and their interpretation. Examples of unstructured data are textual data, imaging data and audio data.

Features (or predictors)

Variables (e.g., patient characteristics or other type of predictors) that are used as input to the AIPA and are used by the AIPA to calculate the probability of having (diagnosis) or developing (prognosis) the individual health outcome in machine learning. In deep learning, the features are identified by the AIPA itself based on the variables in the training data.

Outcome

The target variable(s) of interest, that is/are predicted by the AIPA (e.g., in case of diagnosis, the presence or absence of a specific target condition).

Generalizability

The ability of an AIPA to generalize its operations to other populations or settings, where for example the exclusion of a certain group of patients from the training data leads to the AIPA's results not generalizing to other types or groups of patients. Generalizability should not be confused with bias (see below), and is sometimes confusingly referred to as external validity.

Model scope

Model scope is related to generalizability and indicates the group of people for whom the model produces adequate predictions.

Bias

Bias is generally defined as the presence of the systematic error that leads to distorted or flawed results that hampers the internal validity of a study or analysis. In the case of prediction model and thus also AIPA studies, bias occurs when shortcomings in the design, conduct or analysis could lead to systematically distorted estimates of the AIPA's predictive performance. Bias should not be confused with a lack of generalizability (see above) of a developed AIPA to different populations or settings (which sometimes confusingly is referred to as external validity).

Data drift

The systematic changes in the population, clinical or operational practices over time causing shifts in the data distribution compared to that of the training data used for the development of the AIPA, and may therefore affect the predictive accuracy of the AIPA.

Broad context

Prediction algorithms or models, for diagnostic and prognostic purposes, have a prominent role in healthcare research and practice. The global increase in available digital (health) data, computing power and all-round digitalization of society has raised the academic and commercial interests to explore and exploit the information contained in medical data sources. As a result, on top of the prevailing statistical prediction algorithm methods, methods from machine learning (ML), such as deep learning, and its broader field of Artificial Intelligence (AI) have seen a rapid increase in popularity over the past decade. The latter two research fields have had a strong focus on development of data-driven algorithms in the last decade. While the opportunities of ML and AI in healthcare are undeniable, the growth of complex data-driven prediction algorithms requires careful quality, applicability and validity assessment, before they are used and disseminated in healthcare daily practice. Perhaps as a consequence of the uncertainty in the field regarding this type of quality assessment for ML and AI in healthcare, relatively few applications have been implemented to date in routine clinical use. Accordingly, the (often) established guidelines relating to the earlier developmental stages stand in stark contrast with the meagre and rather exploratory literature regarding the later stages of implementation.

We define the term *AI prediction algorithm* (AIPA) as used in this report as follows: a data-driven algorithm that provides patient-level predictions of a certain outcome (e.g., a certain patient condition), given certain input (e.g., certain patient characteristics or, genetic markers, medical images, other type of predictors). Such prediction algorithms can include supervised and unsupervised algorithms. We use the abbreviation AIPA in the remainder of this report.

Scope of this report

This report provides a non-exhaustive overview of currently available guidelines and quality-criteria and is to be used as a starting point to familiarize or reacquaint oneself with this literature. It discusses guidelines and quality-criteria relevant for construction, testing, software development, evaluating and scaling of AIPA, that form a part of medical (software) devices[1]. These may for instance be used for determining diagnosis (e.g., whether at the moment of prediction an individual has a certain outcome or condition), prognosis (e.g., estimating the risk that an individual develops a certain outcome in the future) or patient monitoring (e.g., during treatment or other patient management situations). We hereby distinguish six phases in AIPA construction that should be considered before embarking on AIPA development. These phases were chosen to provide structure but note that small differences in their ordering and contents could occur depending on each individual developmental lifecycle.

- **Phase 1. Preparation and checking of the data:** Phase 1 consists of the preparation and checking of the data to facilitate proper AIPA development (phase 2) and AIPA validation (phase 3). These data checks and preparations could include (but are not restricted to): assessment of the representativeness of the data for the target population(s), determining the required amount of data needed for AIPA development and/or validation, verifying the legal bases for the different data processing steps, de-identification of personal data, recoding or combining data variables, and investigation and handling (e.g., imputing) of missing values. The operations performed in this phase are performed on both the so-called development data (used in phase 2) and the test or validation data (used in phase 3).
- **Phase 2. Development of the prediction algorithm:** The aim of phase 2 is to model the relation between the predictive input variables (predictors / features) and the health outcome of interest, via a mathematical formula or algorithm. The aim is to develop an AIPA that can be used to predict the outcome for new individuals. Modeling methods to formulate this mathematical relation include prevailing statistical regression techniques, support vector machines, neural networks, and many others.
- **Phase 3. Validation of the predictive algorithm:** Phase 3 consists of testing (validating) how well or accurate the developed AIPA from phase 2, indeed predicts the outcome in new individuals whose data were not used in the model development (internal or external validation data), and to quantify the AIPA's predictive performance. This is done by application of the developed AIPA in the data of new individuals that were not used for the model development and are representative of the intended setting or context in which the developed AIPA is to be used, to calculate the individual's predicted risks of the outcome and by comparing those risks to the true observed outcomes for those individuals.
- **Phase 4. Development of the software application:** The aim of phase 4 is to make an inventory of the norms and recommendations regarding the programming, design, usage and support of the digital packaging of the AIPA. In this context, aspects of the software that relate to AIPA specifically are among others the testing and continuous monitoring of the prediction algorithm (also see phase 6) and aiding the transparency of the decision making process. Also, aspects that relate to the use of the software in general are relevant, such as the user-interaction design. In addition, an inventory is made regarding the industrial and legal frameworks relevant to the software.

[1] A medical device as described in article 2 paragraph 1 of the Medical Device Regulation (MDR).

- **Phase 5. Impact assessment of the AIPA with software:** Phase 5 is concerned with the assessment of the impact of the usage of the AIPA and software on clinical practice and on patient outcomes, where initial cost-effectiveness analysis of the use of the AIPA can also be addressed in this phase. Clinical impact can be defined in several ways and could for example be achieved by supporting the healthcare professional in the clinical decision making process, or by supporting the shared decision making process of care providers. In this phase we present methods from the literature for studying the clinical impact of the AIPA, like cluster-randomized trials where patients in the 'AIPA group' and a 'clinical practice as usual group' are compared with regards to process and patient outcomes.
- **Phase 6. Implementation and use in daily practice:** The goal of the final phase is to provide an overview of the norms and recommendations regarding the implementation of the AIPA in routine medical care. This involves a wide array of technical, scientific and organisational aspects, like bettering the understanding of the end user through education. Moreover, as the underlying population and/or care processes can evolve over time, the AIPA will have to be monitored and regularly recalibrated. Furthermore, this phase examines which norms and recommendations apply to the scalability of the AIPA. This may also involve a more extensive cost-effectiveness analysis and an agreement on the structural financing among other things.

Methods

A systematic inventory of guidance and quality criteria on AIPA that are part of a medical device was performed through two complementary activities: i) the consultation of experts and ii) systematic search of the literature in May 2020.

i) Consultation of experts.

We consulted experts for input on (industry) norms, standards and scientific literature. The interview results were used in conjunction with the outcome of the literature search (see below). The following experts were contacted for relevant and important publications on the topic:

Name	Affiliation	Expertise	Phases
Maarten de Rijke	University of Amsterdam	Artificial intelligence	1-3
Evangelos Kanoulas	University of Amsterdam	Machine learning and statistics	1-3
Floor van Leeuwen	Quantib	Medical device regulation	1-3
Daniel Oberski	Utrecht University	Machine learning and statistics	1-3
Wiro Niessen	Erasmus Medical Center	Medical image processing	1-3
Giovanni Cina	Pacmed	Artificial intelligence	4-6
Rene Aarnink	Philips	Artificial intelligence	4-6
Patrick Jones	Philips	Medical device regulation	4-6
Bart-Jan Verhoeff	Dokter.ai	Clinical software development	4-6
Bart Geerts	Healthplus.ai & Amsterdam UMC	Clinical AI implementation	4-6
Egge van der Poel	Erasmus Medical Center	Personalized healthcare	4-6
Stephan Romeijn	Leiden University Medical Center	Clinical AI implementation	4-6
Martijn Bauer	Leiden University Medical Center	Clinical AI implementation	4-6
André Dekker	Maastricht University	Clinical Data Science	1-6

ii) Systematic literature search.

We conducted a systematic search of the literature in several online scholarly databases (Google Scholar, PubMed and Web of Science). The search strings can be found in the appendix.

In addition, we searched the websites of relevant healthcare institutions based on their interest and expertise of AIPAs in the context of medical devices, on their reporting language (English), and on the recommendation from the above-mentioned experts or our own team members. The websites of the following institutions were searched for relevant guidelines and quality-criteria defined for the construction and testing of AIPAs, that form part of medical (software) devices:

- NHS^{x} (www.nhsx.nhs.uk)
- NICE (www.nice.org.uk)
- FDA (www.fda.gov)
- Health Canada (www.canada.ca/en/health-canada)
- European Commission (https://ec.europa.eu)
- IMDRF (www.imdrf.org)
- ISO (www.iso.org)
- IAIS (www.iais.fraunhofer.de)

Finally, we inspected references cited in found publications (cross references).

Selection of relevant publications

Publications were included if one of the following criteria were met:

- The publication provided guidance on one or more of the six phases of the AIPA implementation cycle (formulated above): inside or outside the healthcare domain.
- The publication discussed challenges on incorporating AI (or ML) in the healthcare domain: either from a technical point of view, or from a regulatory point of view.

Publications that were too specific, and focused on *one* specific example AIPA, a specific AIPA development or validation method, or on a specific type of input variables or outcome, were not included.

This resulted in 143 relevant publications, which can be found in the attached overview of publications (at https://bit.ly/pubs-aipa-guidance). Publications include guidance on many levels, and include institutional documents, scientific articles and widely used scientific textbooks or chapters.

Data extraction and analysis

For all included publications, we extracted the following information:

- To what level of detail the publication includes guidance with regard to one of the six phases of AIPA development (as formulated above).
- To what level of detail the publication includes guidance with regard to development of policy or regulation for the AIPA implementation cycle.
- Whether or not the publication looks at AIPAs in general or has a focus on healthcare applications.
- Publication metadata (incl. source and year of publication).

We start by briefly summarizing the main institutional guidance documents, which provide mostly high-level guidance. Second, we provide an overview of relevant issues for each phase, its related quality instruments and norms as they were mentioned in the literature and the key publications for the selected six stages of the AIPA implementation cycle. Lastly, we provide advice for creating a field norm.

Brief summary of key institutional guidance documents

In recent years, many national and international institutions have started developing guidance documents for AI, including for AIPA construction. Five important principles covered broadly (Jobin et al., 2019) in the found institutional guidance documents are:

- Transparency (e.g., is it clear how an AIPA comes to its prediction?)
- Fairness (e.g., does the AIPA make unbiased, non-discriminating, predictions for all relevant targeted groups?)
- Non-maleficence (e.g., is usage of the AIPA expected to actually benefit the targeted groups?)
- Responsibility (e.g., who carries responsibility for the AIPA's predictions?)
- Privacy (e.g., is an individual's privacy respected during the introduction and usage of an AIPA?)

Most institutional documents structure their guidance by a set of principles, rather than by the consecutive phases of AIPA construction (defined above). Although most principles are relevant in all six previously described phases of AIPA construction, we summarize for each principle to what phases of AIPA construction it is – in our view – most relevant. For example, privacy is important in data preparation for AIPA development (phase 1), but also when incorporating AIPA in daily practice (phases 5 and 6), as the legal basis required for data processing may differ between various settings, and privacy regulation may not only apply to collected data, but also to inferred data (EPRS, 2020). Assessing fairness and non-maleficence (ensuring robustness of the AIPA) can be considered part of the data preparation (phase 1), AIPA development (phase 2), validation (phase 3), software development (phase 4) and impact assessment (phase 5). Transparency is strongly related to AIPA development (phase 2), and the software implementation (phase 4), as these two phases influence to what degree the predictions by the AIPA can be explained and how they are presented to the user. Responsibility should be considered from the beginning, and at the latest when incorporating the AIPA in daily practice (phase 5 and 6).

We now very briefly summarize in our view the most important 9 (out of 35) institutional guidance documents that were found. Four were specifically addressing AIPA in the healthcare domain and 3 were documents providing generic guidance on AIPAs. All 7 documents discuss the principles mentioned above which we will therefore not repeat.

On AI in Healthcare:

- *The United Kingdom National Health Service* (NHS, 2019) have developed a code of conduct for data-driven healthcare technology, that sets out the behaviors expected from those deploying and using data-driven AIPA technologies, to ensure that all those in this chain abide by the ethical principles developed by the Nuffield Council on Bioethics. Adherence to the code of conduct is primarily relevant in phase 1 and 3.
- The *United Kingdom's National Institute for Health and Care Excellence* (NICE, 2019) has developed standards for the evidence that should be available for digital health technologies (of which AIPAs form part) to describe their value in the UK healthcare system, including evidence of effectiveness in the intended use(s) (phase 2 and 3) and also of economic impact in relation to the financial risks. NICE reports specifically that the standards are suitable only for static AIPA, and not adaptive ones, for which they refer to the NHS's code of conduct for data-driven healthcare technology.

- The *International Medical Device Regulators Forum (IMDRF, 2019)* had a meeting in September 2019 with multiple speakers presenting on standards for AI in medical devices. The slides of this meeting are made available online and are an informative resource on regulations and initiatives in various countries.
- *The U.S. Food and Drug Administration (U.S. FDA, 2019)* has proposed a draft regulatory framework for AI applications in software as a medical device (SaMD). An important aspect raised in their report is whether the AIPA always predicts the same results by the same input features (called "**locked/static**"), or whether the AIPA updates itself over time during use in daily practice (called "**continuous/adaptive**"). The FDA stresses the importance of extra quality management after approval for continuous AIPA. The notion of locked versus continuous updating is primarily relevant to phase 2 and 3.

On AI in General:

- *The European Commission* (EC, 2019) has recently published ethics guidelines on trustworthy AI, aiming to provide a framework across various application domains, among which the healthcare domain. Their report stresses the importance of the requirements for general trustworthy AI (including the principles above), which are relevant across all phases. Like the FDA, they stress the importance of continuous monitoring of quality in daily practice for continuous AIPA (in items 100-102 of the report). A checklist of guiding questions on the above-mentioned principles starts on page 32 of the report.
- The *Fraunhofer Institute for Intelligent Analysis and Information Systems* (FIAIS, 2019) has released a white paper (in German) on general certification of AI. In their report they stress the importance of fairness, transparency, autonomy, control, data protection, safety, security and reliability. These are aspects that are relevant for all phases.
- An interesting exploration of the above-mentioned principles has also been published (in Dutch) by the *Authority for the Financial Markets* (AFM, 2019). This publication describes the potential effects of incorporating AI applications in the Dutch insurance sector, and highlights points of attention that require further in-depth exploration.
- The *Asilomar Artificial Intelligence Principles* (Future of Life Institute, 2017) were published in 2017 in conjunction with the Asilomar Conference. The aim is to provide guidance on creating beneficial intelligence. Amongst others, some of the considered topics are responsibility, human control, personal privacy, liberty-privacy, failure transparency, judicial transparency, safety, risks, shared benefit, shared prosperity, common good and non-subversion.
- In collaboration with NGOs, academics, specialists, and policy developers, the Montréal University published the *Montréal Declaration for a Responsible Development of Artificial Intelligence* in 2018. In their report, they stress the importance of well-being, respect for autonomy, protection of privacy and intimacy, solidarity, democratic participation, equity, diversity inclusion, prudence, responsibility and sustainable development.

Key scientific publications

We highlight six sets of key scientific publications: for the preparation of the data, for AIPA development, for AIPA validation, for AIPA software development, for AIPA impact assessment, and for AIPA implementation and scalability. We selected key publications that, together:

- Provide detailed guidance covering a wide range of relevant aspects.
- Are likely to reflect a broad consensus (e.g., are widely cited, or published by relevant institutions or recognized authors in the field).

In each of the following sections, we first summarize important aspects to consider per phase, and then discuss the selected key publications and the respective motivations for inclusion. If we believe guidance is missing in a certain area, we will highlight this separately.

Phase 1. Preparation of the data

An important requirement to yield justifiable inferences from your data is valid data selection and preparation. To fully appreciate any study results, various aspects have to be addressed. Most of these are common, and thus thoroughly discussed, in statistics and clinical epidemiology (Moons et al., 2015; Steyerberg et al., 2019; Riley et al., 2019).We consider:

1. *Source of data (selection), i.e.,*
    a. ***Representativeness of the data.*** The representativeness of the data collected should govern subsequent analysis. The data and the approach (design) that was used to collect the data must fit the study purpose (e.g. diagnostic or prognostic), target population(s), and type of prediction model (e.g., support vector machine, neural network or other). This aspect cannot be understated (Steyerberg et al., 2019).
    b. ***Quality of data.*** The quality of data may influence the validity (absence of bias) of the AIPA. Quality considerations include the richness of the participant data and the quality of the measurement instruments that were used to document the (predictor and outcome) data in the individuals that form the data set. Other components of data quality are:
        i. ***Variety of data.*** The type of data regulates which statistical or data science methods used for AIPA development and validation are applicable. Structured data, i.e., tabular data, can directly be used by all prevailing prediction modeling techniques whilst unstructured data, i.e., raw text and images, primarily need preprocessing. Both structured and unstructured data can be part of an AI/ML prediction model.
        ii. ***Volume of data.*** The size of the available data steers the AIPA development and validation methods that can be used, and can invoke important practical considerations for data storage and data sharing.
2. *Curation of data (preparation), i.e.,*
    a. ***Handling of Missing data.*** The occurrence of incomplete (predictor or outcome) data is exceedingly common, especially when it comes to routine care data which in turn is frequently used in AIPA development. Missing data undeniably impacts study results; its prevention, management and recurrent nature should therefore be a fundamental concern (Donders et al., 2006; Van Buuren, 2018; Sterne et al., 2009; Vergouwe et al., 2010).
    b. ***Standards.*** Sharing information between electronic health records (EHR) of multiple clinical centers requires the use of an interoperable infrastructure of health information (i.e., standards). Clinical terminology lacks the ability to serve all purposes for which clinical data is used. Hence, standards are a requirement for generating useful health information for AIPA development and validation (Bowman et al., 2005; Cases et al., 2013; Wilkinson et al., 2016; Rosenbloom et al., 2017).
    c. ***Harmonization, linkage and privacy.*** Referring to the process of integrating various sources of data, well-founded data harmonization is progressively more essential when more data sources and types (e.g., text, images and figures) become available. Properly harmonizing all available data – both predictors and outcomes - allows for more extensive and useful inferences, while adhering to the GDPR and other privacy guidelines of data use (Kubben et al., 2019; EPRS, 2020), by taking measures such as data de-identification (USDHHS, 2012).

Key publications

We found various (unordered) key publications with guidance relevant for proper data preparation for AIPA development (phase 2) and validation (phase 3).

Key	K-DAT-HARR (v1 2005, v2 **2015**)	**Chapters:** 3, and 4
Title	Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis	
Authors	Frank E. Harrell	
Reason for Inclusion	An important textbook in the field of statistical prediction model development. With regard to data preparation, it offers important guidance on dealing with missing data and consideration of data volume (sample size).	

Key	K-DAT-STEY (v1. 2009, v2. **2019**)	**Chapters:** 3, and 7
Title	Clinical prediction models	
Authors	Ewout Steyerberg	
Reason for Inclusion	K-DAT-STEY is partially overlapping with K-DAT-HARR but has a stronger focus on development of prediction models in the *clinical* domain. It provides an overview of multiple aspects of data preparation, such as promoting data quality for clinical prediction models through study design, sample size considerations, measurement error and dealing with missing values. A number of clinical examples are included.	

Key	K-DAT-RILEY (**2019**)	**Chapters:** 13 and 14
Title	Prognosis Research in Health Care: Concepts, Methods, and Impact	
Authors	Richard D. Riley, Danielle van der Windt, Peter Croft, and Karel G.M. Moons	
Reason for Inclusion	Further extended guidance with regard to study design, and the origin of the collected data is covered by K-DAT-RILEY. Overall, it has a focus on prognostic outcomes. Beside more classic study designs (prospective, retrospective or case-control designs), it provides guidance on harmonization of individual participant data, and using data from electronic health records.	

Key	K-DAT-MOON (**2015**)	**Items:** 4-9, 13
Title	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration	

Authors	Karel G.M. Moons, Douglas G. Altman, Johannes B. Reitsma, John P.A. Ioannidis, Petra Macaskill, Ewout W. Steyerberg, Andrew J. Vickers, David F. Ransohoff, Gary S. Collins
Reason for Inclusion	K-DAT-MOON provides a concise overview of important aspects mentioned above. It described the established TRIPOD reporting standard (a checklist of 22 items) for clinical prediction model development and validation, including multiple items that cover relevant aspects for data preparation (items 4-9, and 13).

Key	K-DAT-WILL (**2020**)	**Pages**: all
Title	Preparing Medical Imaging Data for Machine Learning	
Authors	Martin J. Willemink , Wojciech A. Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R. Folio, Ronald M. Summers, Daniel L. Rubin, Matthew P. Lungren	
Reason for Inclusion	Although much of the provided guidance is data-type independent, the guidance focuses on using regression techniques for AIPA development, notably using tabular data. Depending on the data type (images, video's, textual domains, audio, etc.), and the specific application of the AIPA, different specialized data preparation may be required. As many recently developed AIPA include imaging data in combination with machine learning methods (Litjens et al., 2017; Miotto, 2018), we include K-DAT-WILL. K-DAT-WILL describes general considerations when preparing imaging data for AIPA development, including de-identification (to better preserve patient privacy), representativeness of image data, possible down sampling of images (lowering resolution for both computational efficiency and prediction quality reasons), and also mentions data standards like the Digital Imaging and Communications in Medicine (DICOM) format.	

Key	K-DAT-BENS (**2018**)	**Chapter**: 1 and 2
Title	Principles of Health Interoperability	
Authors	Tim Benson, and Grahame Grieve	
Reason for Inclusion	Beside the DICOM format, there exist many more data standards in healthcare, to facilitate interoperability of data for various purposes. We suggest chapters 1 and 2 of K-DAT- BENS, which provides a nice overview of various well-known data standards in healthcare, like the International Classification of Diseases (ICD), and the Systematized Nomenclature of Medicine (SNOMED) terminological standard, among others (later chapters cover technical details on specific standards).	

Key	K-DAT-FORT (**2016**)	**Pages:** all
Title	Maelstrom Research guidelines for rigorous retrospective data harmonization	

Authors	Isabel Fortier, Parminder Raina, Edwin R Van den Heuvel, Lauren E Griffith, Camille Craig, Matilda Saliba, Dany Doiron, Ronald P Stolk, Bartha M Knoppers, Vincent Ferretti, Peter Granda, Paul Burton
Reason for Inclusion	Finally, we provide K-DAT-FORT, which has established guidance on how to perform high quality clinical data harmonization, based on many existing efforts in the literature, and by consulting multiple experts in the field. The paper also provides information on common pitfalls. Data harmonization was not yet discussed in the preceding 4 other key publications.

Phase 2. Development of the AIPA

AIPA development consists of determining a function that models the relation between the predictive input variables (*predictors* or *features*) and the to be predicted outcome variables (e.g., patient outcomes) by means of a mathematical formula or algorithm. The development process is an interplay between effective use of the available data, careful incorporation of domain knowledge, and further AIPA requirements (e.g., explainability). Important aspects to consider in this phase are:

1. *Relevant outcome:* The predicted outcome is what drives all steps of AIPA construction. Carefully establishing a patient relevant outcome to be predicted, ideally chosen with input from healthcare professionals and patients, that can be measured accurately, is one of the first steps of AIPA development.
2. *Predictive performance measures*: Predictive performance measures reflect how well the AIPA predicts the outcome, and preferably have a direct link to clinical relevance (Matheny et al., 2020). In most cases, not only the model's discriminative performance (how well is the algorithm capable of distinguishing between individuals with and without the outcome?), but also calibration (how well the algorithm's predictions agree with the actual observed outcomes) is critical (Van Calster et al., 2019). Relevant evaluation measures, and their desired performance are generally determined at an early stage of AIPA development.
3. *Informative input variables*: A crucial aspect is to select the *predictors*, or *features* that are informative for predicting the outcome (e.g., based on current literature, clinical knowledge, availability, and cost of the data collection). Including uninformative features or missing informative features can reduce AIPA quality, accuracy (validity) and generalizability.
4. *Model specification and training*: The prediction function specifies the operations that are performed on the input values to come to the predicted outcome. Generally, these operations involve parameters (or *coefficients*) that need to be estimated using data (a process called algorithm or model *fitting* or *training*). For complex data types (imaging, audio or text), often the prediction function is layered (or *deep*). Sometimes complex prediction functions yield better predictions, but can also be maintenance-heavy, poorly generalizable, and may require separate solutions to make decisions explainable, in contrast to simpler algorithms.
5. *Overfitting*: If an AIPA is adapted too much to the development/training data to such a degree that it's predictions no longer generalize well to future individuals, the AIPA is overfitted. The smaller the sample size, the more input variables are being considered, and the more model specification is influenced by data, the higher the risk of algorithm overfitting to the data at hand, making the algorithm hard to generalize and apply to new individuals.

Key Publications

We found multiple (unordered) key publications providing general guidance relevant for development of AIPA.

Key	K-DEV-HARR (v1 2005, v2 **2015**)	**Chapters:** 2, 4, 7, 9, 10-21
Title	Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis	
Authors	Frank E. Harrell	
Reason for Inclusion	For AIPA development, K-DEV-HARR provides guidance on model specification and overfitting through detailed explanations of important regression approaches to general prediction model development, including model specification and training. Different types of outcomes are discussed, including binary outcomes, continuous outcomes, ordinal outcomes, and survival models.	

Key	K-DEV-STEY (v1 2009, v2 **2019**)	**Chapters**: 4-6, 9-14
Title	Clinical Prediction Models	
Authors	Ewout W. Steyerberg	
Reason for Inclusion	K-DEV-STEY focuses on *clinical* prediction models, and includes detailed guidance on, e.g., defining the relevant outcomes, selecting the informative input variables, and particularly on technical aspects like prediction modeling, specification, overfitting and predictive performance measurement. It covers many statistical regression techniques for various types of outcomes, and gives many empirical examples.	

Key	K-DEV-HAST (v1 2009, v12 **2017**)	**Chapters:** all
Title	The Elements of Statistical Learning: Data Mining, Inference, and Prediction	
Authors	Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman	
Reason for Inclusion	As in K-DEV-STEY, and K-DEV-HARR the focus lies on regression modeling techniques, we include K-DEV-HAST, possibly the most widely[2] used (technical) textbook on general statistical learning techniques, covering a very wide range of methods, mostly for model specification and training, including ML and AI approaches. It is written by important researchers in the field of statistical learning and includes frequently used approaches like unsupervised learning and clustering, tree-based methods, ensembles, and graphical models.	

[2] Google scholar citations: 49.199 (on 04-06-2020)

Key	K-DEV-MOON (**2015**)	**Items:** 6, 7, 10, 14, and 15
Title	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration	
Authors	Karel G.M. Moons, Douglas G. Altman, Johannes B. Reitsma, John P.A. Ioannidis, Petra Macaskill, Ewout W. Steyerberg, Andrew J. Vickers, David F. Ransohoff, Gary S. Collins	
Reason for Inclusion	Beside the AIPA development aspects mentioned above, reporting of choices made incorporating each aspect is crucial. K-DEV-MOON is a reporting standard for clinical prediction model development and validation, providing an overview of important items that should be included when reporting on AIPA development.	

Key	K-DEV-GOOD (**2016**)	**Chapter:** 11
Title	Deep Learning	
Authors	Ian Goodfellow and Yoshua Bengio and Aaron Courville	
Reason for Inclusion	In the recent years, there have been important new developments in the field of neural networks. As the above-mentioned key publications do not cover many of the recent developments, we suggest K-DEV-GOOD, a widely used[3] textbook by leading researchers in the field, as complementary key publication in this respect. K-DEV-GOOD does not focus on clinical AIPA specifically. But it gives practical guidance on developing AIPA with *deep* prediction functions (deep learning), including guidance on dataset size, choosing networks structures, overfitting and model estimation.	

Key	K-DEV-MOLN (**2020**)	**Chapters:** 2 , 4-7
Title	Interpretable Machine Learning: A Guide for Making Black Box Models Explainable	
Authors	Christoph Molnar	
Reason for Inclusion	Deep learning AIPA are sometimes considered as “black box” models, due to their complex prediction functions. As explainability of AIPA predictions is almost always an important requirement in healthcare practice, we suggest K-DEV-MOLN as a key publication. K-DEV-MOLN describes how and to what degree AIPA, and their predictions can be explained. It provides a clear, and recent overview of types of AIPA explanations (Chapter 2), and model-specific and model-agnostic methods (Chapters 4-7) to provide such explanations for various AIPA. The book is focused on (but not limited to) tabular data.	

[3] Google scholar citation: 15.970 (on 04-06-2020)

Phase 3. Validation of the AIPA

The goal of validation is to apply a developed AIPA to the data of new individuals whose data were not used in the algorithm or model development (validation set), and to quantify the algorithm's predictive performance. Note that preparation of the data (from phase 1) is also required to properly prepare the validation data. Validation can be done retrospectively and prospectively, although the latter are seldom performed. Validation includes the following aspects (although more aspects may be relevant):

1. *Predictive performance assessment*: An AIPA can be validated by making predictions for new individuals in a validation dataset (a set disjoint from the development data) using the algorithm, and consecutive assessment of the algorithm's discrimination and calibration. The type of population used as the validation set (see also phase 1) determines the degree of validity: individuals can be used from the same population as the development data (split-sample internal validation), or from a different population (external validation: e.g., data from a different time period of the same institute, from a different institute or from different countries). External validation is widely considered as required before use into medical practice, to ensure a form of robustness and generalizability of the developed algorithm (also see phase 5). Sometimes artificially created difficult settings (called adversarial examples) are constructed to further assess robustness (also see phase 4).
2. *Early clinical impact assessment:* Beside raw predictive performance assessment, a first preliminary clinical and cost-effectiveness assessment can be performed, preceding phase 4 and 5, to estimate the expected clinical impact of the AIPA in the target setting, which is (in part) based on clinical judgement of the relative value of benefits and harms associated with the AIPA, and alternative markers, and tests (so-called net benefit analysis) (Vickers et al., 2006; 2016).

Key Publications

We found multiple (unordered) key publications with guidance relevant for validation of AIPA.

Key	K-VAL-STEY (v1. 2009, v2. **2019**)	**Chapters:** 15 and 17
Title	Clinical Prediction Models	
Authors	Ewout W. Steyerberg	
Reason for Inclusion	K-VAL-STEY, a textbook on statistical clinical prediction model development and validation, provides extensive detailed recommendations in Chapters 15 and 17 on validation of clinical prediction models, in terms of validation procedures as well as recommendations on validation sample size, and discussion on predictive performance measures.	

Key	K-VAL-ALTM (**2000**)	**Pages:** all
Title	What do we mean by validating a prognostic model?	
Authors	Douglas G. Altman and Patrick Royston	
Reason for Inclusion	A more detailed explanation of *prognostic* AIPA validation is discussed by K-VAL-ALTM. Important aspects in data preparation for the validation set are discussed. It provides an extensive explanation on validation of prognostic models, separating clinical and statistical validation, discussing AIPA transportability, stressing the importance of pre-specifying adequate performance thresholds, study design to obtain validation data, and inspection of coefficients. Additionally, it discusses validation of several exemplary models from the literature in case studies.	

Key	K-VAL-MATH (**2019**)	**Pages:** 170, 131-135
Title	Artificial Intelligence in Health Care: The hope, the Hype, the Promise, the Peril	
Authors	Michael Mathey, Sonoo Thadaney Israni, Mahnoor Ahmed, Danielle Whicher, *Editors*	
Reason for Inclusion	K-VAL-MATH, a National Academy of Medicine publication on considerations and key issues when incorporating AI models into healthcare, provides an overview discussion on predictive AI model evaluation in healthcare. Like K-VAL-STEY it stresses the importance of discrimination and calibration, connects issues and terminology from both AI and prevailing statistical prediction modeling, and indicates that measures and methods used in prevailing prediction modeling may be suitable for evaluating predictive AI algorithms in many (if not most) cases. Additionally, it stresses the importance of using (train and test) data that is representative of the target population, and takes into account patients of different socioeconomic, cultural, and ethnic backgrounds is important to impede or detect potential bias.	

Key	K-VAL-SOKO (**2009**)	**Pages:** all
Title	A systematic analysis of performance measures for classification tasks	
Authors	Marina Sokolova, Guy Lapalme	
Reason for Inclusion	As K-VAL-STEY and K-VAL-MATH both focus on time-to-event data and binary outcomes, we also include K-VAL-SOKO, a widely cited publication providing an overview of discrimination performance measures for classification beyond the binary case (e.g., multi-class classification and hierarchical classification). Although binary outcomes are probably most frequently modeled in healthcare, more complex label-structures exist and may become more prevalent in the future (e.g., ICD-coding already requires evaluation with a large hierarchical label structure).	

Key	K-VAL-MOON (**2015**)	**Items:** 4-9,13 (validation data), 11, 12, 16, 19
Title	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration	
Authors	Karel G.M. Moons, Douglas G. Altman, Johannes B. Reitsma, John P.A. Ioannidis, Petra Macaskill, Ewout W. Steyerberg, Andrew J. Vickers, David F. Ransohoff, Gary S. Collins	
Reason for Inclusion	K-VAL-MOON, a reporting standard for clinical prediction model development and validation, provides guidance on proper reporting of AIPA validation. Beside providing items that ensure clear reporting on the quality of the validation dataset (items 4-9, and 13), it requires reporting of performance measurement on both development and validation data, among other aspects (items 11, 12, 16, and 19).	

Key	K-VAL-CADE (**2012**)	**Pages:** 15-22
Title	Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data	
Authors	Food and Drug Administration	
Reason for Inclusion	As final key publication for AIPA validation, we put forward the FDA's guidance on computer-assisted detection devices in radiology, which includes ML-based AIPA, and covers important topics like: selection of representative patients, measuring outcome, comparing predictive performance measures (with existing devices), among other aspects. Although focused on radiology, its guidance may be applicable to medical domains outside radiology where AIPAs play a role. We categorize this one under validation as it provides detailed guidance on stand-alone validation of AIPA.	

Phase 4. Software development for AIPA

Software development is the gateway to evaluating and implementing a properly developed (phase 2) and validated (phase 3) AIPA into daily medical practice. It is the packaging that connects the AIPA to databases and makes it easily accessible to the end user. Various aspects play an important role in ensuring the quality of AIPA specific software development. The following aspects are considered:

1. *An inventory of related industrial and legal frameworks.*
    a. ***Legal frameworks.*** The European Union's *Medical Device Regulation* (MDR, 2017) enforces quality demands on a software product which is classified as a medical device. The MDR risk classification is based on endpoints, such as risks of the software application, adverse effects, and analytical and clinical performance. A software product is classified as a medical device requiring conformity assessment if it falls in risk class IIa (software that has a therapeutic or diagnostic goal) or higher. The risk class will increase proportionally to the potential negative effects or risks to the patient related to the use of the 'AIPA-software' product. This happens regardless of whether underlying prediction algorithms used in the software can be classified as AI or not.
    b. ***Industry frameworks.*** Some of the most internationally recognized standards are produced by the IEEE, ISO and IEC (Holt, 1996). Both IEEE and ISO are working on AI specific software engineering standards, like the IEEE P7000 standards (ISO, 2020; OCEANIS, 2020), but at time of writing neither has delivered their finished products.
2. *Design.* The layered architecture pattern is the most common form of software architecture (Richards, 2015). It consists of several layers, where the database, persistence, business and presentation layer are considered standard. The AIPA is embedded in the business layer which performs business logic on the data it receives from the database and persistence layer.
    a. ***Database, persistence & business layers.*** Recommendations on the design for the software comprising these first three layers can largely be found in the industry standards as discussed above. However, little is known about whether these recommendations extend to and sufficiently cover AIPA specific software architecture. To ensure good quality design, the following stages could be taken into account when planning and developing the AIPA-software (POOD, 2002): i) user requirement specification, ii) software design, iii) coding, iv) testing (see phase 4 *Testing*) and v) maintenance.
    b. ***Presentation layer.*** The presentation layer or front-end through which the end user will interact with the AIPA should fit the care provider's workflow (Magrabi et al., 2019) (also see phase 6 *Interaction design and user satisfaction*). For example, front-end design forms the perfect medium to make hard-to-interpret probabilities accessible through visualisation etc. Moreover, the front-end design is the ideal platform to address the issues of transparency and explainability (Buruk, Ekmekci & Arda 2020; European Commission, 2019). A mindful design can open the black box by visualizing the results and providing insight into the workings of the AIPA for the end user (see phase 5 *Explainability*), thereby fitting the requirement mentioned in the *General Data Protection Regulation* (GDPR, 2016).
3. *Testing.*
    a. ***Conventional software testing.*** Conventional software quality testing focuses on accessibility, reliability, flexibility, integration, performance, robustness, security, etc. (Magrabi et al., 2019; Tao, Gao, & Wang, 2019). Some of these quality parameters, like performance, can and will be addressed in

the validation steps of phase 3 and phase 4, and in the impact assessment of phase 5 (in consultation with a medical professional). Generally speaking, (AI) software is tested using black-box testing techniques, meaning functionality is tested without the knowledge of its internal operating mechanism (Tao et al., 2019).

b. ***AIPA software testing: parameters.*** AIPA software applications form unique challenges as they are, generally speaking, more complex, highly reliant on (changeable) data, and above all non-deterministic in their outputs (Tao et al., 2019). This means that the 'golden path' testing procedure described above likely will not provide sufficient quality control. To this extent, additional quality parameters, like system stability and data quality can also be taken into account (Tao et al., 2019).

c. ***AIPA software testing: context.*** AIPAs might be sensitive to different contexts, in which case it is recommended to test AIPA software in changing contexts and environments (Tao et al., 2019).

d. ***Continuous monitoring.*** Some AIPAs have the potential to continuously learn and develop over their lifetime, which enforces the need for continuous monitoring and to some extent testing (Buruk et al., 2020; Magrabi et al., 2019). During software development, the digital infrastructure for this type of continuous quality control has to be built (Buruk et al., 2020; Magrabi et al., 2019). For more information on continuous monitoring see *Post-market surveillance* in phase 6.

4. *Security.*

a. ***Robustness.*** The robustness of the AIPA system should ensure the security of the end user's data by protecting it from hardware vulnerabilities and "operational or system interacting agents" (Buruk et al., 2020; European Commission, 2019). The developers should take the appropriate precautions in terms of security, safety and privacy early on to prevent unwanted outcomes (Buruk et al., 2020; Future of Life, 2017; MDRDAI, 2018).

b. ***Software attacks.*** Software attacks specifically directed at software with underlying AIPA come in two forms: i) altering the composition of the training data and ii) training the AIPA in such a way that it elicits a prespecified response at runtime (Hahn, Ebner-Priemer, & Meyer-Lindenberg, 2019). The latter issue is more problematic in that it requires specific immunization of the AIPA through so-called adversarial attacks. The first issue should be addressed by broadening the scope and increasing the amount of the training data. Both issues should be tested before release and continuously monitored afterwards (Hahn et al., 2019). The GDPR (2016) and MDR (2017) are leading in embedding these security precautions during the AIPA software development.

Key Publications

We found multiple (unordered) key publications with guidance relevant for software development of AIPA.

Key	K-SOFT-BUR (**2020**)	**Pages:** all
Title	A critical perspective on guidelines for responsible and trustworthy artificial intelligence.	
Authors	Buruk, B., Ekmekci, P. E., & Arda, B.	
Reason for Inclusion	An article on three major ethical guidelines produced to date for responsible AI. It pertains to software development in that it discusses themes of (software) security, quality monitoring and explainability of AIPAs.	

Key	K-SOFT-HAHN (**2019**)	**Pages:** all
Title	Transparent artificial intelligence – A conceptual framework for evaluating AI-based clinical decision support systems	
Authors	Hahn, T., Ebner-Priemer, U., & Meyer-Lindenberg, A.	
Reason for Inclusion	A discussion paper suggesting a framework with which to evaluate AIPA quality, clinical utility and security based on the concept of AI transparency. It gives insights into the specific issues regarding AIPA software security.	

Key	K-SOFT-MAG (**2019**)	**Pages:** all
Title	Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications	
Authors	Magrabi, F., Ammenwerth, E., McNair, J. B., De Keizer, N. F., Hyppönen, H., Nykänen, P., Rigby, M., Scott, P. J., Vehko, T., Wong, Z. S. Y., & Georgiou, A.	
Reason for Inclusion	A position paper from the IMIA Technology Assessment & Quality Development in Health Informatics Working Group and the EFMI Working Group for Assessment of Health Information Systems. It discusses key considerations for evaluating artificial intelligence and the challenges and practical implications of AI design, development, selection, use, and ongoing surveillance. In terms of software development, it discusses testing frameworks, front-end design related to explainability and quality monitoring.	

Key	K-SOFT-TAO (**2019**)	**Pages:** all
Title	Testing and Quality Validation for AI Software-Perspectives, Issues, and Practices	

Authors	Tao, C. Q., Gao, J., & Wang, T. X.
Reason for Inclusion	An article concerned with the testing and quality validation for AI software and how it differs from ‘regular’ software.

Phase 5. Impact assessment

Impact assessment is required to determine the value of AIPA use for medical practice, including for the targeted patients and even for society at large (cost-effectiveness with a societal perspective). It can to a certain extent draw on the extensive literature regarding impact assessment for other health technology innovations, in particular of prevailing types of medical prediction models. However, some AIPA specific aspects should also be taken into account. We consider:

1. *Prospective validation.* Prospective validation is closely related to the more retrospective validation as is described in phase 3, with the difference that prospective validation is performed on a dataset not available at the time of developing or training the AIPA. To achieve proper prospective assessment of validity (absence of bias) and generalisation, it is recommended that researchers use large, heterogeneous, (multi-centre) datasets curated from other institutions than those providing the training data (Altman & Royston, 2000; Altman, Vergouwe, & Royston, 2009; Cearns et al., 2019; Kelly et al., 2019; Moons et al., 2012; Steyerberg et al., 2013). This is ideally done by independent investigators (Vollmer et al., 2020). The provision of obtaining sufficient information on the input features and the to be predicted outcomes by the researchers is crucial to the success of the prospective validation (Magrabi et al., 2019; Vollmer et al., 2020; Wiens et al., 2019).
2. *Bias.* Even though the problem of bias will be addressed as much as possible during AIPA development and retrospective validation (see phase 2 and 3), the bias could persist when introduced in the new clinical setting (Kelly et al., 2019).
    a. ***Creating awareness.*** To increase awareness of bias, it is recommended to compile a diverse team of regulatory scholars, social scientists and ethicists to help the AI experts navigate and mitigate bias throughout the life cycle of the AIPA (Wiens et al., 2019; Moons et al., 2019).
    b. ***Heterogeneity and generalizability (sometimes referred to as external validity).*** Bias as a result of underrepresentation of certain subgroups within the target population will lead to a lower predictive accuracy of these subgroups (Debray et al., 2015; Vollmer et al., 2020; Wiens et al., 2019; Moons et al., 2019). This type of heterogeneity with a lack of generalizability as a result, can be mitigated by paying special attention to the inclusion of these subgroups during data collection and training (phase 1, 2 and 3).
3. *Explainability.* Explainable AI refers to the ability of the prediction algorithm to justify its predictions to a certain extent (Cearns et al., 2019; Magrabi et al., 2019). Sufficient explainability can help mitigate various types of adverse effects (including bias and security breaches, see phase 4 *Security* and phase 5 *Bias*) and it is crucial to the trustworthiness of the AIPA (Cearns et al., 2019; Kelly et al., 2019; Magrabi et al., 2019; Vollmer et al., 2020). However, explainability of AIPAs could signal a level of causal inference it simply cannot provide as generally the underlying associations will rely on correlation rather than causation (van Hartskamp, Consoli, Verhaegh, Petkovic, & van de Stolpe, 2019). Hence, explainability should be accompanied by the thorough education of end users (see phase 6) and a continuous vigilance to ensure the correct use of the AIPA. During impact assessment, the level and presentation of explainability can be assessed in close collaboration with the end user (also see phase 6 *Interaction design and user satisfaction*).
4. *Clinical impact by incremental utility.*
    a. ***Clinical benchmarking.*** Incremental utility indicates the added value of an AIPA with respect to current practice in terms of the decision making process, patient outcomes and costs-effectiveness

(Altman et al., 2009; Cearns et al., 2019; Kappen et al., 2018; Moons et al., 2009; Moons et al., 2012; Steyerberg et al., 2013). It assesses whether the AIPA is operationally meaningful (Cearns et al., 2019; Holt, 2005; Magrabi et al., 2019). It is important to note here that high predictive accuracy does not necessarily translate into clinical impact or effectiveness (Cearns et al., 2019; Kelly et al., 2019). Therefore, a first step in assessing the impact on patient outcomes, costs-effectiveness and decision making would be to compare the effectiveness of using the AIPA *relative to* the current state of the art clinical practice in which the AIPA is not used (Holt, 2005).

b. ***Randomised trial.*** The randomized clinical trial (RCT) is often named as the gold standard to assess this type of impact (Cearns et al., 2019; Kelly et al., 2019; Moons et al., 2009; Moons et al., 2012; Steyerberg et al., 2013; Wiens et al., 2019;). However, randomization at the patient or care provider's level may not always be feasible due to the drastic disruption in workflow required (Moons et al., 2012; Steyerberg et al., 2013; Wiens et al., 2019). An alternative comparative effectiveness study could be a pre-post study with seasonality corrections or a – preferably randomised cluster - stepped-wedge design as this type of trial introduces the changes gradually (Kappen et al., 2018; Moons et al., 2012; Wiens et al., 2019).

5. *Risks and unintended consequences.*
    a. ***Level of confidence.*** The level of confidence surrounding the prediction of risk for an individual can vary greatly between or within individuals. It is therefore highly encouraged to always give some level of model confidence with the algorithmic prediction of individual risks, which should emerge from phase 2 and 3 (Kelly et al., 2019; Magrabi et al., 2019; Wiens et al., 2019).
    b. ***Unintended consequences.*** Unintended consequences of the use of an AIPA in daily practice could range from a security breach to unanticipated human behaviour. They will have to be studied in an RCT in addition to the impact on decision making, patient outcomes and costs-effectiveness (see phase 5 *Clinical impact by incremental utility*). Qualitative approaches could complement the risk assessment by exposing concerns of the end users and other stakeholders associated with the performance of the AIPA (Wiens et al., 2019).

Key Publications

We found multiple (unordered) key publications with guidance relevant for impact assessment of AIPA.

Key	K-IA-CEA (**2019**)	**Pages:** all
Title	Recommendations and future directions for supervised machine learning in psychiatry	
Authors	Cearns, M., Hahn, T., & Baune, B. T.	
Reason for Inclusion	An extensive review article providing guidelines on how the systematically evaluate the claims, maturity, and clinical readiness of a project. It is applied to the field of psychiatry, but easily generalizable to other medical domains. The topics relating to impact assessment are prospective validation, model scope, bias detection, explainability, clinical trials and reporting.	

Key	K-IA-HOLT (**2005**)	**Pages:** all
Title	Clinical benchmarking for the validation of AI medical diagnostic classifiers	
Authors	Holt, G.	
Reason for Inclusion	A letter to the editor providing a short and very clear outline of how to approach clinical benchmarking for the validation of an AIPA.	

Key	K-IA-KEL (**2019**)	**Pages:** all
Title	Key challenges for delivering clinical impact with artificial intelligence	
Authors	Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D.	
Reason for Inclusion	An article discussing key challenges for the translation of AIPAs in healthcare, including machine learning specific issues, logistical difficulties and barriers to adoption as well as necessary sociocultural or pathway changes. It pertains to all the impact assessment topics mentioned above.	

Key	K-IA-MAG (**2019**)	**Pages:** all
Title	Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications	
Authors	Magrabi, F., Ammenwerth, E., McNair, J. B., De Keizer, N. F., Hyppönen, H., Nykänen, P., Rigby, M., Scott, P. J., Vehko, T., Wong, Z. S. Y., & Georgiou, A.	
Reason for Inclusion	A position paper from the IMIA Technology Assessment & Quality Development in Health Informatics Working Group and the EFMI Working Group for Assessment of Health Information Systems. It discusses key considerations for evaluating artificial	

	intelligence and the challenges and practical implications of AI design, development, selection, use, and ongoing surveillance. It discusses all impact assessment topics mentioned above.

Key	K-IA-MOONS (**2012**)	**Pages:** all
Title	Risk prediction models: II. External validation, model updating, and impact assessment.	
Authors	Moons, K. G. M., Kengne, A. P., Grobbee, D. E., Royston, P., Vergouwe, Y., Altman, D. G., & Woodward, M.	
Reason for Inclusion	Paper focusses on external validation studies, updating of models and investigating impact of models on clinical decision-making and patient outcomes. The paper is set within the cardiovascular domain, but its contents can easily be generalized to other domains. The paper addresses generalizability and impact assessment through many different study design.	

Key	K-IA-VOL (**2020**)	**Pages:** all
Title	Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness	
Authors	Vollmer, S., Mateen, B. A., Bohner, G., Kiraly, F. J., Ghani, R., Jonsson, P., Cumbers, S., Jonas, A., McAllister, K. S. L., Myles, P., Grainger, D., Birse, M., Branson, R., Moons, K. G. M., Collins, G. S., Ioannidis, J. P. A., Holmes, C., & Hemingway, H.	
Reason for Inclusion	An article providing a framework for (amongst others) patients, clinicians and policy makers to critically appraise where new findings may deliver patient benefit. In terms of impact assessment, it regards prospective validation, bias detection, explainable AI, risks and unintended consequences and reporting of the results.	

Key	K-IA-WIENS (**2019**)	**Pages:** all
Title	Do no harm: a roadmap for responsible machine learning for health care	
Authors	Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A.	
Reason for Inclusion	An article providing a framework and guidelines for accelerating the translation of machine-learning-based interventions in health care. It pertains to all the impact assessment topics mentioned above.	

Phase 6. Clinical implementation and scalability

The implementation of AIPA into daily medical care involves a wide range of technical, scientific and organizational aspects. For implementation and scalability into daily practice we consider the following aspects:

1. *Post-market surveillance.* Post-market surveillance refers to the continuous evaluation of the data (to expose changes in the population and/or discrimination), the use of the AIPA and the AIPA itself and possible periodical recalibration of the AIPA after routine clinical adoption (Kelly et al., 2019; Magrabi et al., 2019; Vollmer et al., 2020; Wiens et al., 2019). It is a requirement under the European Union's *Medical Device Regulation* (MDR, 2017) and should be balanced with the privacy of personal health data (see *Data privacy*). If left unchecked, data shift will lead to deteriorating algorithmic predictive performance and subsequent patient outcomes (Vollmer et al., 2020). Developers should consider and integrate the mechanisms required for this type of monitoring early on in the development process (Buruk et al., 2020; Future of Life, 2017; MDRDAI, 2018; Vollmer et al., 2020) (see phase 4 *Testing*). In addition, developers could make use of the implementation and evaluation guidelines mentioned by Magrabi et al. (2019) in shaping the monitoring requirements (Canada Health Infoway Benefits Evaluation Indicators Technical Report version 2.0. 2012.; DeLone & McLean, 1992; DeLone & McLean, 2003; Hyppönen et al., 2013; Schloemer & Schröder-Bäck, 2018).
2. *Data privacy.* Data privacy will have to be ensured at a larger scale in this phase and continuously monitored (Magrabi et al., 2019).
    a. ***Autonomy.*** Personal health data should stay confidential and AI technology applied to the data should under no circumstances restrict people's freedom (Buruk et al., 2020; European Commission, 2019; Future of Life, 2017; MDRDAI, 2018). The GDPR (2016) can be consulted as the leading source on data protection in the European Union. For more information on how the GDPR affects AIPA development, see the EPRS rapport (2020).
    b. ***Consent.*** From an ethical perspective the developers should respect the threshold of consent for the use of personal data (Floridi et al., 2020). This level of consent may vary per application and is relevant for all phases of development.
3. *Interaction design and user satisfaction.* The technical aspects of the front-end design (presentation layer) were listed in the software development phase (*Design*). During implementation, the design can be optimized through user feedback (Floridi et al., 2020; Vollmer et al., 2020). The interaction design is closely related to explainability (phase 5) as it can provide insight into the workings of the AIPA and through this increase the end user's trust in the system (Vollmer et al., 2020). The end goal should always be to make the design fit clinical workflow (Magrabi et al., 2019).
4. *Education of end users.* The educational needs of the end user will have to be addressed to ensure a safe and successful implementation (Kelly et al., 2019; Magrabi et al., 2019). The literature has put forth the idea of an easily accessible AI curriculum for medical students and healthcare practitioners that could focus on the critical appraisal, adoption and usage of AI tools in clinical practice (Kelly et al., 2019). For more information on the approach taken in the Netherlands, see the covenant medical technology and assessment framework provided by the Dutch Ministry of Health, Welfare and Sport.
5. *Certification.* Distribution and commercialization of medical software require compliance with region specific standards for health, safety, and environmental protection (Wiens et al., 2019). It is a hard requirement for the

scaling of the product, but ideally addressed earlier in the AIPA developmental cycle (see phase 4 *An inventory of related industrial and legal frameworks*). Note that, for continuous learning prediction algorithms, regular review of regulatory compliance is necessary throughout the lifetime of the AIPA, as its performance *will* change over time (Wiens et al., 2019)[4].

6. *Cost-effectiveness.* Cost-effectiveness of AI clinical care in comparison with usual care is needed to guarantee a sustainable widespread adoption of the technology and to inform the decision maker's appraisal of the technology (Vollmer et al., 2020) (also see phase 5 *Clinical impact by incremental utility*).
7. *Distribution financial benefits and structural financing.* Structural financing indicates the financial arrangements between the technology provider, health care insurers and/or the government as they are necessary for maintaining model safety and wide operational success (Wiens et al., 2019). The distribution of financial benefits should also be discussed at this stage (Vollmer et al., 2020).

[4] To this end, the U.S. Food & Drug Administration has been working on the *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)* (2019).

Key Publications

We found multiple (unordered) key publications with guidance relevant for the implementation and scalability of AIPA.

Key	K-IMP&SC-BUR (**2020**)	**Pages:** all
Title	A critical perspective on guidelines for responsible and trustworthy artificial intelligence.	
Authors	Buruk, B., Ekmekci, P. E., & Arda, B.	
Reason for Inclusion	An article on three major ethical guidelines produced to date for responsible AI. It pertains to implementation and scalability in that it discusses themes of post-market surveillance and data privacy.	

Key	K-IMP&SC-FLOR (**2020**)	**Pages:** all
Title	How to Design AI for Social Good: Seven Essential Factors	
Authors	Floridi, L., Cowls, J., King, T. C., & Taddeo, M.	
Reason for Inclusion	An article considering seven factors essential for AI for social good initiatives and their corresponding best practices that may serve as preliminary guidelines. With regards to implementation and scalability it discusses data privacy and interaction design.	

Key	K-IMP&SC-KEL (**2019**)	**Pages:** all
Title	Key challenges for delivering clinical impact with artificial intelligence	
Authors	Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D.	
Reason for Inclusion	An article discussing key challenges for the translation of AIPAs in healthcare, including machine learning specific issues, logistical difficulties and barriers to adoption as well as necessary sociocultural or pathway changes. The topics related to implementation and scalability are post-market surveillance and education of end users.	

Key	K-IMP&SC-MAG (**2019**)	**Pages:** all
Title	Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications	
Authors	Magrabi, F., Ammenwerth, E., McNair, J. B., De Keizer, N. F., Hyppönen, H., Nykänen, P., Rigby, M., Scott, P. J., Vehko, T., Wong, Z. S. Y., & Georgiou, A.	

Reason for Inclusion	A position paper from the IMIA Technology Assessment & Quality Development in Health Informatics Working Group and the EFMI Working Group for Assessment of Health Information Systems. It discusses key considerations for evaluating artificial intelligence and the challenges and practical implications of AI design, development, selection, use, and ongoing surveillance. In terms of implementation, it discusses all respective topics mentioned above (post-market surveillance, data privacy, interaction design and user satisfaction, education) and in terms of scalability it discusses certification.

Key	K-IMP&SC-VOL (**2020**)	**Pages:** all
Title	Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness	
Authors	Vollmer, S., Mateen, B. A., Bohner, G., Kiraly, F. J., Ghani, R., Jonsson, P., Cumbers, S., Jonas, A., McAllister, K. S. L., Myles, P., Grainger, D., Birse, M., Branson, R., Moons, K. G. M., Collins, G. S., Ioannidis, J. P. A., Holmes, C., & Hemingway, H.	
Reason for Inclusion	An article providing a framework for (amongst others) patients, clinicians and policy makers to critically appraise where new findings may deliver patient benefit. With regards to implementation and scalability, it discusses post-market surveillance, interaction design and user satisfaction, cost-effectiveness and distribution of financial benefits.	

Key	K-IMP&SC-WIENS (**2020**)	**Pages:** all
Title	Do no harm: a roadmap for responsible machine learning for health care	
Authors	Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A.	
Reason for Inclusion	An article providing a framework and guidelines for accelerating the translation of machine-learning-based interventions in health care. It pertains to implementation and scalability in that it focusses on post-market surveillance, certification and distribution of financial benefits and structural financing.	

Ongoing initiatives

Besides the provided institutional documents, and proposed key publications, several initiatives have been announced, and are currently under development. Two important initiatives are:

- The developers of TRIPOD, a widely used reporting standard for clinical prediction models, have announced (Collins et al., 2019) the development of TRIPOD-ML, a new reporting standard for studies aimed at developing and validating clinical prediction models based on AI and ML, and TRIPOD-CLUSTER, an extension of TRIPOD for reporting prediction model studies conducted in big routine care or registry datasets.
- CONSORT-AI and SPIRIT-AI are preparing international AI-specific extensions to the CONSORT and SPIRIT statements with a specific focus on clinical trials in which the intervention includes an AIPA.
- ISO has set up a committee (ISO, 2020) to develop international standards for **general** AI (on terminology, but also engineering aspects). This initiative will continue for at least three more years (possibly more).
- The IEEE launched the IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems (in 2016), concerning, amongst other things, transparency, algorithmic bias considerations and data privacy (OCEANIS, 2020).

Discussion

This report provides a non-exhaustive overview of currently available guidelines and quality-criteria relevant for the implementation cycle of AIPAs that are part of medical (software) devices and to be used for diagnosis, prognosis, therapy monitoring and other medical purposes. We provided several key publications that can form an introduction into the guidelines and quality-criteria of the six phases comprising the AIPA implementation cycle: (1) preparation of the data, (2) development of the AIPA, (3) validation of the AIPA, (4) software development of the AIPA (5) impact assessment of the AIPA and (6) implementation and scaling of the AIPA. Again, we would like to remind the reader that these phases – and the order in which they are presented – are not set in stone and variations can occur depending on each specific situation. Nevertheless, they form the building blocks for AIPA construction and provide a useful structure for the assembled literature.

Due to the vast variety in possible medical settings (e.g., general practice vs. hospital care), data types (e.g., tabular data vs. combinations of unstructured data), methods to develop AIPA (e.g., logistic regression vs. deep learning), and scale of intended use (e.g., internal vs. widely distributed) it is important to realize that applicability of the guidance in the key publications, as well as their completeness may depend on the use case. Additional application-specific, data-type-specific, method-specific, and scale specific criteria may be required to make a proper AIPA quality assessment. Also, some AIPA methods and uses are relatively new and require further research and regulation (e.g., development and validation of continuous AIPA, that update over time in production, AIPA immunization). We therefore also refer to the attached[5] longer list of references that can be reviewed for their applicability to the six phases.

The reviewers noted a stark contrast between the literature available on the first three and last three phases. The first three phases have received considerable attention, which was apparent in the number of book chapters relating to these phases. Several aspects of the latter three phases have been relatively underexposed, especially in terms of AIPA specific literature and use cases. This can perhaps be explained by the fact that relatively few AIPA applications have been implemented at the time of writing this report. As a result, one of the five ethical principles mentioned at the beginning of the section *Brief Summary of Key Institutional Guidance Documents* (Jobin et al., 2019), responsibility, was barely discussed in the literature.

[5] https://bit.ly/pubs-aipa-guidance

Creating a field norm for healthcare AI

This document is intended as a basis for the creation of a field norm regarding medical devices with AIPAs. To this end, the authors examined the quality instruments and norms reported in the current literature pertaining to all six phases of AIPA development. The reported findings can inform and guide the field norm expert panel on each respective phase. In creating a field norm, we recommend the following based on our observations in the literature:

1. *Build on existing frameworks*. Multiple legal and industrial frameworks already in place are – to varying extent – relevant to the six phases of AIPA development. We have highlighted them where applicable and advise the panel to create a thorough understanding of the frameworks in place, prior to the formation of additional guidance for AIPA medical devices. Not all frameworks will be specifically developed for AIPAs but will nonetheless provide a useful and necessary starting point, such as clinical guidelines and quality guideline regulations like the MDR.
2. *Adhere to ethical principles*. In the creation of a new field norm medical ethical principles should be taken into account. Five broadly covered principles in other guidance documents are: transparency, fairness, non-maleficence, responsibility and privacy.
3. *Mind the gap.* Not all phases of the AIPA implementation cycle have been equally well studied and documented. Where topics like overfitting and impact assessment in clinical trials are reflected in a large literature base, others such as updates and upgrades of AIPA software and corresponding certification, software architecture as it pertains to AIPA, long-term effects of AIPAs in healthcare, domain adaptation, effects on the workforce and responsibility have hardly been touched upon in the literature at all. In terms of the creation of a field norm, the panel should take these potential ‘gaps’ in the literature into account.
4. *Advance and assess the continuous research effort*. Currently no real consensus exists on what should be the appropriate course of action, even for certain AIPA aspects that enjoy an extensive literature base. Examples that come to mind are explainable AI, fairness in AI, uncertainty estimation of AIPAs and causal inference. Continuous research and experimentation are needed to establish best practices through which a field norm can be informed and updated accordingly.

References

Altman, D. G., & Royston, P. (2000). What do we mean by validating a prognostic model?. *Statistics in medicine*, 19(4), 453-473.

Altman, D. G., Vergouwe, Y., Royston, P., & Moons, K. G. (2009). Prognosis and prognostic research: validating a prognostic model. *Bmj*, *338*, b605. doi:10.1136/bmj.b605

Authority for the Financial Markets (2019). Artificiële Intelligentie in de verzekeringssector: een verkenning. Retrieved from: https://www.afm.nl/~/profmedia/files/rapporten/2019/afm-dnb-verkenning-ai-verzekeringssector.pdf

Benson, T., & Grieve, G. (2016). *Principles of health interoperability: SNOMED CT, HL7 and FHIR*. Springer.

Bowman, S. E. (2005). Coordination of SNOMED-CT and ICD-10: getting the most out of electronic health record systems. *Coordination of SNOMED-CT and ICD-10: Getting the Most out of Electronic Health Record Systems/AHIMA, American Health Information Management Association*.

Buruk, B., Ekmekci, P. E., & Arda, B. (2020). A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Medicine Health Care and Philosophy*. doi:10.1007/s11019020-09948-1

Canada Health Infoway Benefits Evaluation Indicators Technical Report version 2.0. 2012. Retrieved from:https://www.infowayinforoute.ca/en/component/edocman/resources/reports/benefitsevaluation/184-benefits-evaluation-indicators-technical-report

Cases, M., Furlong, L. I., Albanell, J., Altman, R. B., Bellazzi, R., Boyer, S., ... & Gea, J. (2013). Improving data and knowledge management to better integrate health care and research. *Journal of internal medicine*, 274(4), 321.

Cearns, M., Hahn, T., & Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Transl Psychiatry*, *9*(1), 271. doi:10.1038/s41398-019-0607-2

Collins, G. S., & Moons, K. G. (2019). Reporting of artificial intelligence prediction models. *The Lancet*, 393(10181), 1577-1579.

Debray, T. P., Vergouwe, Y., Koffijberg, H., Nieboer, D., Steyerberg, E. W., & Moons, K. G. (2015). A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol*, *68*(3), 279-289. doi:10.1016/j.jclinepi.2014.06.018

DeLone, W. H., & McLean, E. R. (1992). Information systems success: The quest for the dependentvariable. Information Systems Research, 3, 60-95.

DeLone, W. H., & McLean, E. R. (2003). The DeLone and McLean model of information systems success: A ten year update. Journal of Management Information Systems, 19, 9-30.

National Health Service. (2019). Code of conduct for data-driven health and care technology. Retrieved 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

Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59(10), 1087-1091.

European Commission. (2019). *EGTAI: the ethics guidelines for trustworthy artificial intelligence*.Retrieved from: https://ec.europa.eu/futurium/en/ai-alliance-consultation

European Parliament; Council of the European Union (27 April 2016). *Regulation (EU) 2016/679 ofthe 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)*. Retrieved from: https://eur-lex.europa.eu/eli/reg/2016/679/oj

European Parliament; Council of the European Union (5 April 2017). *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 repeating Council Directives 90/385/EEC and 93/42/EEC. Retrieved from: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1592131978075&uri=CELEX:320170745

European Parliamentary Research Service. (2020). The impact of the General Data Protection Regulation (GDPR) on artificial intelligence. Retrieved from: https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2020)641530

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to Design AI for Social Good: Seven Essential Factors. *Sci Eng Ethics*. doi:10.1007/s11948-020-00213-5

Fortier, I., Raina, P., Van den Heuvel, E. R., Griffith, L. E., Craig, C., Saliba, M., ... & Granda, P. (2017). Maelstrom Research guidelines for rigorous retrospective data harmonization. *International journal of epidemiology*, 46(1), 103-105.

Fraunhofer Institute for Intelligent Analysis and Information Systems (2019). Building trust in artificial intelligence. Retrieved from: https://www.iais.fraunhofer.de/en/press/press-release-190702.html

Future of Life. (2017). *AAIP: Asilomar AI Principles*. Retrieved from:https://futureoflife.org/ai-principles/

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Hahn, T., Ebner-Priemer, U., & Meyer-Lindenberg, A. (2019). Transparent artificial intelligence – A conceptual framework for evaluating AI-based clinical decision support systems. *OSF Preprints*. Retrieved from: https://doi.org/10.31219/osf.io/uzehj

Harrell Jr, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Holt, J. D. (1996). Practical issues in the design of AI software. Engineering Applications of Artificial Intelligence, 9(4), 429-437. doi:10.1016/0952-1976(96)00029-2

Holt, G. (2005). Clinical benchmarking for the validation of AI medical diagnostic classifiers. *Artif Intell Med, 35*(3), 259-260. doi:10.1016/j.artmed.2005.10.001

Hyppönen, H., Faxvaag, A., Gilstad, H., Hardardottir, G. A., Jerlvall, L., Kangas, M., . . . Vimarlund, V. (2013). Nordic eHealth Indicators: Organisation of research, first results and the plan for the future. Retrieved from http://norden.diva-portal.org/smash/get/diva2:700970/FULLTEXT01.pdf

International Medical Device Regulators Forum, (2019). Meetings. Retrieved May, 2020, from: http://www.imdrf.org/meetings/meetings.asp

ISO. (2020). *ISO/IEC JTC 1/SC 42 - Artificial Intelligence*. Retrieved from: https://www.iso.org/committee/6794475.html

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

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

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Med, 17*(1), 195. doi:10.1186/s12916-019-1426-2

Kubben, P., Dumontier, M., & Dekker, A. (2019). Fundamentals of clinical data science. *Springer Nature*.Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42, 60-88.

Magrabi, F., Ammenwerth, E., McNair, J. B., De Keizer, N. F., Hyppönen, H., Nykänen, P., . . . Georgiou, A. (2019). Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications. *Yearb Med Inform, 28*(1), 128-134. doi:10.1055/s-00391677903

Matheny, M., Israni, S. T., Ahmed, M., & Whicher, D. (2020). Artificial intelligence in health care: The hope, the hype, the promise, the peril. *Natl Acad Med*, 94-97.

MDRDAI. (2018). Montréal declaration for a responsible development of artificial intelligence. Retrieved from: https://www.montrealdeclaration-responsibleai.com/

Ministerie van Volksgezondheid, Welzijn en Sport. *Convenant medische technologie.* Retrieved from: https://www.igj.nl/zorgsectoren/medische-technologie/toezicht-op-veilig-gebruik/convenant

Ministerie van Volksgezondheid, Welzijn en Sport. *Uitgebreide versie van het toetsingskader 'Inzet van ehealth door zorgaanbieders'*. Retrieved from: https://www.igj.nl/documenten/toetsingskaders/2018/11/15/uitgebreide-versie-van-het-toetsingskader-inzet-van-e-health-door-zorgaanbieders

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6), 1236-1246.

Moons, K. G., Altman, D. G., Vergouwe, Y., & Royston, P. (2009). Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *Bmj, 338*, b606. doi:10.1136/bmj.b606

Moons, K. G., Kengne, A. P., Grobbee, D. E., Royston, P., Vergouwe, Y., Altman, D. G., & Woodward, M. (2012). Risk prediction models: II. External validation, model updating, and impact assessment. Heart, 98(9), 691-698. doi:10.1136/heartjnl-2011-301247

Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., Macaskill, P., Steyerberg, E. W., ... & Collins, G. S. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of internal medicine*, 162(1), W1-W73.

National Institute for Health and Care Excellence. (2019). Evidence standards framework for digital health technologies. Retrieved from: https://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/digital-evidence-standards-framework.pdf

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447-453.

OCEANIS. (2020). *IEEE P7000™ Projects.* Retrieved from: https://ethicsstandards.org/p7000/.

Principles of Object-Oriented Development. (2002). VU. Retrieved from: https://www.cs.vu.nl/~eliens/oop/0.html.

Richards, M. (2015). *Software Architecture Patterns*. O'Reilly Media, Inc. Retrieved from: https://www.oreilly.com/library/view/software-architecture-patterns/9781491971437/

Riley, R. D., van der Windt, D., Croft, P., & Moons, K. G. (Eds.). (2019). *Prognosis research in healthcare: concepts, methods, and impact*. Oxford University Press.

Rosenbloom, S. T., Carroll, R. J., Warner, J. L., Matheny, M. E., & Denny, J. C. (2017). Representing knowledge consistently across health systems. *Yearbook of medical informatics*, 26(01), 139-147.

Schloemer, T., & Schröder-Bäck, P. (2018). Criteria for evaluating transferability of health interventions: a systematic review and thematic synthesis. Implementation Science, 13(88).

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4), 427-437.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338.

Steyerberg, E. W., Moons, K. G., van der Windt, D. A., Hayden, J. A., Perel, P., Schroter, S., . . . Altman, D. G. (2013). Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med, 10*(2), e1001381. doi:10.1371/journal.pmed.1001381

Steyerberg, E. W. (2019). *Clinical prediction models*. Springer International Publishing.

Tao, C. Q., Gao, J., & Wang, T. X. (2019). Testing and Quality Validation for AI Software-Perspectives, Issues, and Practices. *Ieee Access, 7*, 120164-120175. doi:10.1109/access.2019.2937107

U.S. Food & Drug Administration. (2020). *Computer-assisted detection devices applied to radiology images and radiology device data—Premarket notification [510 (k)] submissions*. Retrieved from: https://www.fda.gov/media/77635/download

U.S. Food & Drug Administration. (2019). *Proposed Regulatory Framework for Modifications to Artificial Intelligence/ Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)*. Retrieved from: https://www.fda.gov/media/122535/download

U.S. Department of Health and Human Services. (2012). *Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*. Retrieved from: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: the Achilles heel of predictive analytics. *BMC medicine*, 17(1), 1-7.

Van Hartskamp, M., Consoli, S., Verhaegh, W., Petkovic, M., & Van de Stolpe, A. (2019). Artificial Intelligence in Clinical Health Care Applications: Viewpoint. Interact J Med Res, 8(2), e12100. doi:10.2196/12100

Vergouwe, Y., Royston, P., Moons, K. G., & Altman, D. G. (2010). Development and validation of a prediction model with missing predictor data: a practical approach. *Journal of clinical epidemiology*, 63(2), 205-214.

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6), 565-574.

Vickers, A. J., Van Calster, B., & Steyerberg, E. W. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *bmj*, 352, i6.

Villongco, C., & Khan, F. (2020). "Sorry I Didn't Hear You." The Ethics of Voice Computing and AI in High Risk Mental Health Populations. *AJOB Neurosci, 11*(2), 105-112. doi:10.1080/21507740.2020.1740355

Vollmer, S., Mateen, B. A., Bohner, G., Kiraly, F. J., Ghani, R., Jonsson, P., . . . Hemingway, H. (2020). Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *Bmj-British Medical Journal, 368*. doi:10.1136/bmj.l6927

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., . . . Goldenberg, A. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nat Med, 25*(9), 1337-1340. doi:10.1038/s41591-019-0548-6

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1), 1-9.

Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., ... & Lungren, M. P. (2020). Preparing medical imaging data for machine learning. *Radiology*, 295(1), 4-15.

Acknowledgements

This work was supported by the Dutch Ministry of Health, Welfare and Sport.

APPENDIX

Search strings

Google scholar search for general guidance

("Artificial Intelligence" OR "Machine Learning" OR "Deep Learning" OR "Prediction Model" OR "Supervised Learning" OR "Reinforcement Learning") AND ("Guideline" OR "Guidelines" OR "Guidance" OR "Best Practice" OR "Recommendations" OR "Challenges") AND ("Healthcare" OR "Medicine" OR "Medical" OR "Diagnosis" OR "Prognosis")

Google scholar search for data preparation

("Artificial Intelligence" OR "Machine Learning" OR "Deep Learning" OR "Prediction Model" OR "Supervised Learning" OR "Reinforcement Learning") AND ("Guideline" OR "Guidelines" OR "Guidance" OR "Best Practice" OR "Recommendations" OR "Challenges") AND ("Data" AND ("Preparation" OR "Preparing" OR "Transformation" OR "Standards" OR "Missing" OR "Harmonization" OR "Preprocessing")) OR "Study Design" OR "Outlier Detection"

("Artificial Intelligence" OR "Machine Learning" OR "Deep Learning" OR "Prediction Model" OR "Supervised Learning" OR "Reinforcement Learning") AND ("Guideline" OR "Guidelines" OR "Guidance" OR "Best Practice" OR "Recommendations" OR "Challenges") AND ("Healthcare" OR "Medicine" OR "Medical" OR "Diagnosis" OR "Prognosis") AND ("Data" AND ("Preparation" OR "Preparing" OR "Transformation" OR "Standards" OR "Missing" OR "Harmonization" OR "Preprocessing")) OR "Study Design" OR "Outlier Detection"

Google scholar search for AIPA development

("Artificial Intelligence" OR "Machine Learning" OR "Deep Learning" OR "Prediction Model" OR "Supervised Learning" OR "Reinforcement Learning") AND ("Guideline" OR "Guidelines" OR "Guidance" OR "Best Practice" OR "Recommendations" OR "Challenges") AND ("Development" OR "Developing" OR "Methodology" OR "Methods")

("Artificial Intelligence" OR "Machine Learning" OR "Deep Learning" OR "Prediction Model" OR "Supervised Learning" OR "Reinforcement Learning") AND ("Guideline" OR "Guidelines" OR "Guidance" OR "Best Practice" OR "Recommendations" OR "Challenges") AND ("Healthcare" OR "Medicine" OR "Medical" OR "Diagnosis" OR "Prognosis") AND ("Development" OR "Developing" OR "Methodology" OR "Methods")

Google scholar search for AIPA validation

("Artificial Intelligence" OR "Machine Learning" OR "Deep Learning" OR "Prediction Model" OR "Supervised Learning" OR "Reinforcement Learning") AND ("Guideline" OR "Guidelines" OR "Guidance" OR "Best Practice" OR "Recommendations" OR "Challenges") AND ("Evaluation" OR "Validation" OR "Benchmarking" OR "Performance measure" OR "Evaluation measure")

("Artificial Intelligence" OR "Machine Learning" OR "Deep Learning" OR "Prediction Model" OR "Supervised Learning" OR "Reinforcement Learning") AND ("Guideline" OR "Guidelines" OR "Guidance" OR "Best Practice" OR "Recommendations" OR "Challenges") AND ("Healthcare" OR "Medicine" OR "Medical" OR "Diagnosis" OR "Prognosis") AND ("Evaluation" OR "Validation" OR "Benchmarking" OR "Performance measure" OR "Evaluation measure")

PubMed search software development

(("Artificial Intelligence"[ti] OR "Computational Intelligence"[ti] OR "Machine Intelligence"[ti] OR "Computer Reasoning"[ti] OR "Computer Vision Systems"[ti]OR "Computer Vision System"[ti] OR "AI"[ti] OR "kunstmatige intelligentie"[tt] OR"Computer Heuristics"[ti] OR "Expert Systems"[ti] OR "Expert Systems"[ti] OR "Fuzzy

Logic"[ti] OR "Natural Language Processing"[ti] OR "Robotics"[ti] OR "Support Vector Machine "[ti] OR "Support Vector Machines"[ti] OR "machine learning"[ti] OR "deep learning"[ti] OR "supervised machine learning"[ti] OR "unsupervised machine learning"[ti] OR "Natural Language Processing"[ti] OR "Neural Networks"[ti] OR "Neural Network"[ti]) AND ("Quality of Health Care"[tiab] OR"Benchmarking"[tiab] OR "Healthcare Quality"[tiab] OR "Quality Improvement"[tiab] OR "Quality Indicator"[tiab] OR "Quality Indicators"[tiab] OR "Total Quality Management"[tiab] OR “Richtlijn”[tt] OR “Richtlijnen”[tt] OR “Guidelines”[tiab] OR “Guideline”[tiab] OR "Practice Guideline"[tiab] OR "Practice Guidelines"[tiab] OR "quality norm"[tiab] OR "quality norms"[tiab] OR "quality instrument"[tiab] OR "quality instruments"[tiab] OR "kwaliteitsnorm"[tt] OR "kwaliteitsnormen"[tt] OR "kwaliteitsinstrument"[tt] OR "kwaliteitsinstrumenten"[tt] OR "quality of care"[tiab] OR "quality assessment"[tiab] OR "best practice"[tiab]) AND ("software"[tw] OR "software quality assurance"[tw] OR "SQA"[tw] OR "software quality control"[tw] OR "SQC"[tw] OR "software as a medical device"[tw] OR "SaMD"[tw]OR "softwarekwaliteit"[tt] OR "software kwaliteit"[tt] OR "software compliance"[tw]) AND (english[la] OR dutch[la])AND (english[la] OR dutch[la]))

PubMed search impact assessment

(("Artificial Intelligence"[ti] OR "Computational Intelligence"[ti] OR "Machine Intelligence"[ti] OR "Computer Reasoning"[ti] OR "Computer Vision Systems"[ti]OR "Computer Vision System"[ti] OR "AI"[ti] OR "kunstmatige intelligentie"[tt] OR"Computer Heuristics"[ti] OR "Expert Systems"[ti] OR "Expert Systems"[ti] OR "Fuzzy Logic"[ti] OR "Natural Language Processing"[ti] OR "Robotics"[ti] OR "Support Vector Machine "[ti] OR "Support Vector Machines"[ti] OR "machine learning"[ti] OR "deep learning"[ti] OR "supervised machine learning"[ti] OR "unsupervised machine learning"[ti] OR "Natural Language Processing"[ti] OR "Neural Networks"[ti] OR "Neural Network"[ti]) AND ("Quality of Health Care"[tiab] OR"Benchmarking"[tiab] OR "Healthcare Quality"[tiab] OR "Quality Improvement"[tiab] OR "Quality Indicator"[tiab] OR "Quality Indicators"[tiab] OR "Total Quality Management"[tiab] OR “Richtlijn”[tt] OR “Richtlijnen”[tt] OR “Guidelines”[tiab] OR “Guideline”[tiab] OR "Practice Guideline"[tiab] OR "Practice Guidelines"[tiab] OR "quality norm"[tiab] OR "quality norms"[tiab] OR "quality instrument"[tiab] OR "quality instruments"[tiab] OR "kwaliteitsnorm"[tt] OR "kwaliteitsnormen"[tt] OR "kwaliteitsinstrument"[tt] OR "kwaliteitsinstrumenten"[tt] OR "quality of care"[tiab] OR "quality assessment"[tiab] OR "best practice"[tiab]) AND ("Health Impact Assessment"[tw] OR "Outcome Assessment"[tw] OR "Validation Study"[Publication Type] OR "validation"[tw] OR validat*[tw] OR "validatie"[tt] OR "prospective validation"[tw] OR "retrospective validation"[tw] OR "clinical performance"[tw] OR "clinical investigation"[tw] OR "external validation"[tw] OR "RCT"[tw] OR "Randomized Controlled Trial"[Publication Type] OR "Randomized Controlled Trials"[tw] OR "Randomized Clinical Trial"[tw] OR "Randomised Controlled Trials"[tw] OR "Randomised Clinical Trial"[tw] OR "random control trial" OR "Technology Assessment"[tw] OR "HTA"[tw] OR "clinical impact"[tw] OR "klinische impact"[tt] OR "externe validatie"[tt] OR "pilot study"[tw] OR "clinical benefit"[tw] OR "clinical benefits"[tw])AND (english[la] OR dutch[la]))

PubMed search clinical implementation and scalability

(("Artificial Intelligence"[ti] OR "Computational Intelligence"[ti] OR "Machine Intelligence"[ti] OR "Computer Reasoning"[ti] OR "Computer Vision Systems"[ti]OR "Computer Vision System"[ti] OR "AI"[ti] OR "kunstmatige intelligentie"[tt] OR"Computer Heuristics"[ti] OR "Expert Systems"[ti] OR "Expert Systems"[ti] OR "Fuzzy Logic"[ti] OR "Natural Language Processing"[ti] OR "Robotics"[ti] OR "Support Vector Machine "[ti] OR "Support Vector Machines"[ti] OR "machine learning"[ti] OR "deep learning"[ti] OR "supervised machine learning"[ti] OR "unsupervised machine learning"[ti] OR "Natural Language Processing"[ti] OR "Neural Networks"[ti] OR "Neural Network"[ti]) AND ("Quality of Health Care"[tiab] OR"Benchmarking"[tiab] OR "Healthcare Quality"[tiab] OR "Quality Improvement"[tiab] OR "Quality Indicator"[tiab] OR "Quality Indicators"[tiab] OR "Total Quality Management"[tiab] OR “Richtlijn”[tt] OR “Richtlijnen”[tt] OR “Guidelines”[tiab] OR “Guideline”[tiab] OR "Practice Guideline"[tiab] OR "Practice Guidelines"[tiab] OR "quality norm"[tiab] OR "quality norms"[tiab] OR "quality instrument"[tiab] OR "quality instruments"[tiab] OR "kwaliteitsnorm"[tt] OR "kwaliteitsnormen"[tt] OR "kwaliteitsinstrument"[tt] OR "kwaliteitsinstrumenten"[tt] OR "quality of care"[tiab] OR "quality assessment"[tiab] OR "best practice"[tiab]) AND ("Implementation Science"[tw] OR "Implementation"[tw] OR implement*[tw] OR "implementatie"[tt] OR "integration"[tw] OR

integrat*[tw] OR "integratie"[tt] OR "calibration"[tw] OR "kalibratie"[tt] OR "transfer learning"[tw] OR "usability"[tw] OR "patient satisfaction"[tw] OR "interoperability systems"[tw] OR "interoperability system"[tw] OR "interoperability"[tw] OR "gebruiksvriendelijk"[tt] OR "user friendly"[tw] OR "patiënttevredenheid"[tt] OR "ethics"[Subheading] OR "ethical"[tw] OR "ethics"[tw] OR "ethiek"[tt] OR "Jurisprudence"[tw] OR"legislation and jurisprudence"[Subheading] OR "legislation"[tw] OR "legal issues"[tw] OR "legal"[tw] OR "law"[tw] OR "laws"[tw] OR "juridisch"[tt] OR "Diffusion of Innovation"[tw] OR "Application"[tw] OR "Applications"[tw] OR "Dissemination"[tw] OR Disseminat*[tw] OR "real-world performance"[tw] OR "real world performance"[tw])AND (english[la] OR dutch[la]))

Web of Science search software development

(TI=("Artificial Intelligence" OR "AI" OR "kunstmatige intelligentie" OR "machine learning" OR "deep learning" OR "supervised machine learning" OR "unsupervised machine learning" OR "Natural Language Processing" OR "Neural Networks") AND TS=("Quality of Health Care" OR "Benchmarking" OR "Practice Guidelines as Topic" OR "quality norm" OR "quality instrument" OR "kwaliteitsnorm" OR "kwaliteitsinstrument" OR "quality of care" OR "quality assessment" OR "best practice" OR “Richtlijn” OR “Richtlijnen” OR “Guidelines”) AND TS=("Software" OR "software quality assurance" OR "SQA" OR "software quality control" OR "SQC" OR "software as a medical device" OR "SaMD" OR "softwarekwaliteit " OR "software kwaliteit" OR "software compliance")) NOT DT=(meeting abstract) AND la=(english OR dutch)

Web of Science search impact assessment

(TI=("Artificial Intelligence" OR "AI" OR "kunstmatige intelligentie" OR "machine learning" OR "deep learning" OR "supervised machine learning" OR "unsupervised machine learning" OR "Natural Language Processing" OR "Neural Networks") AND TS=("Quality of Health Care" OR "Benchmarking" OR "Practice Guidelines as Topic" OR "quality norm" OR "quality instrument" OR "kwaliteitsnorm" OR "kwaliteitsinstrument" OR "quality of care" OR "quality assessment" OR "best practice" OR “Richtlijn” OR “Richtlijnen” OR “Guidelines” OR “Guideline”) AND TS=("Health Impact Assessment" OR "Outcome Assessment" OR "validation" OR "validatie" OR "prospective validation" OR "retrospective validation" OR "clinical performance" OR "clinical investigation" OR "external validation" OR "RCT" OR "random control trial" OR "Health Technology Assessment" OR "HTA" OR "clinical impact" OR "klinische impact" OR "externe validatie" OR "pilot study" OR "clinical benefit")) NOT DT=(meeting abstract) AND la=(english OR dutch)

Web of Science search implementation and scalability

(TI=("Artificial Intelligence" OR "AI" OR "kunstmatige intelligentie" OR "machine learning" OR "deep learning" OR "supervised machine learning" OR "unsupervised machine learning" OR "Natural Language Processing" OR "Neural Networks") AND TS=("Quality of Health Care" OR "Benchmarking" OR "Practice Guidelines as Topic" OR "quality norm" OR "quality instrument" OR "kwaliteitsnorm" OR "kwaliteitsinstrument" OR "quality of care" OR "quality assessment" OR "best practice" OR “Richtlijn” OR “Richtlijnen” OR “Guidelines” OR “Guideline”) AND TS=("Implementation Science" OR "Implementation" OR "implementatie" OR "integration" OR "integratie" OR "calibration" OR "calibratie" OR "transfer learning" OR "usability" OR "patient satisfaction" OR "interoperability systems" OR "interoperability" OR "gebruiksvriendelijk" OR "patiënttevredenheid" OR "ethical" OR "ethics" OR "ethiek" OR "legal issues" OR "legal" OR "juridisch" OR "Diffusion of Innovation" OR "Application" OR "Dissemination" OR "real-world performance" OR "real world performance")) NOT DT=(meeting abstract) AND la=(english OR dutch)