Guideline for high-quality diagnostic and prognostic applications of AI in healthcare

Version 1.1 16-08-2023

Authors

Maarten van Smeden, Carl Moons, Lotty Hooft (phase 1 through 3)

Ilse Kant, Hine van Os, Niels Chavannes (phase 4 through 6)

On behalf of the working party members on Medical AI

Commissioned by the Ministry of Health, Welfare and Sport

Waardevolle AI

Contents

Acknowledgements 1

Introduction 3

Status of this document 4

Scope 5

Parties involved 6

Comply or explain 7

Phases of development, testing and implementation 9

Phase 0: Preparations for the development process 9

Creation of the guideline 10

References 11

1 Collection and management of the data 12

1.1 Legal prerequisites 13

1.2 Data collection 14

1.2.1 Privacy and traceability 15

1.3 Metadata 16

1.4 Availability of data 16

1.5 Version management and availability of the data management plan 18

1.6 References 19

2 Development of the AIPA 20

2.1 Explanation of intended use 21

1.1.1. Data set(s) and intended use 22

2.2 Analysis and modelling steps 22

2.3 Internal evaluation of the model 22

2.3.1 Internal validation 22

2.3.2 Analysis of potential (negative) impact of the model 23

2.4 Technical Robustness 24

2.5 Size of the data set for development of the AIPA 24

2.6 Logging, availability and version management 25

2.6.1 Logging, reproducibility and replicability 25

2.6.2 Version management and availability of the model 26

2.7 References 27

3 Validation of the AIPA 30

3.1 Evaluation of predictive (statistical) characteristics of the AIPA 31

3.1.1 Target population and context 31

3.1.2 Model performance of the AIPA 32
3.2 Evaluation of medical characteristics and expectations for implementation of the AIPA 33
3.3 Fairness and algorithmic bias 34
3.4 Determining the outcome variable (labelling) 35
3.5 Size of the data set for external validation 35
3.6 Logging, reproducibility and replicability 36
3.7 References 37
4 Development of the required software 39
4.1 Explainability, transparency, design and information 40
4.1.1 Explainability, transparency and design of the AIPA software 40
4.1.2 Information pertaining to the software 41
4.2 Provisions for continuous monitoring 43
4.3 Security 44
4.4 Software testing 44
4.5 References 46
5 Impact assessment of the AIPA in combination with the software 47
5.1 Impact assessment and setting up accompanying study 48
5.1.1 The expected effects 49
5.1.2 Risk assessment 52
5.1.3 Human-machine interaction 52
5.1.4 Comparative study 53
5.2 Health technology assessment 54
5.3 Uncertainty, risks and unexpected outcomes 55
5.3.1 Uncertainty in predictions 55
5.3.2 Unexpected outcomes, vigilance 55
5.4 References 56
6 Implementation and use of the AIPA with software in daily practice 59
6.1 Implementation plan 60
6.2 Monitoring 62
6.2.1 Responsibilities of manufacturer or developing care organisation 62
6.2.2 Responsibilities of the care organisation 63
6.3 Education 65
6.3.1 End-user 65
6.3.2 Care organisation 66
6.4 Rights and Duties 67
6.4.1 Care provider 67

6.4.2 Care organisation 68
6.4.3 Patient, client or citizen 68
6.4.4 Manufacturer or developing care organisation 69
6.5 References 70
Future perspectives 71
Dynamic updates of AIPAs 72
Cost-effectiveness evaluations 72
Sharing data 72
Early multi-disciplinary work 73
Monitoring in practice 73
Education 73
References 75

Acknowledgements

In 2020, the University Medical Center Utrecht (UMCU) and Leiden University Medical Center (LUMC) performed a literature review and wrote a report in which they provided an overview of the available national and international guidelines and criteria for the development, validation, evaluation and implementation of *Artificial Intelligence Prediction Algorithm* (AIPA) in the medical sector, including the public health sector. Based on this literature review and the subsequent working parties, the UMCU and LUMC drafted a guideline of requirements and criteria for the development, validation, evaluation and implementation of an AIPA in the medical sector.

The efforts of many stakeholders resulted in the creation of this guideline. In particular, we wish to thank *Rosalie van Oosterom* and *Pieter Boone* from the Ministry of Health, Welfare and Sport for organising and coordinating the creation of this guideline, *Roy Tomeij* for chairing the working parties and *Rachel Peeters* for her support during and after the working party meetings. We wish to thank all the action team members and the working party members for their active participation, commitment and engagement in the working parties that have resulted in this guideline. In addition, we would like to thank all reviewers and the participants in the field test for their numerous feedback comments and KPMG for the coordination, which have resulted in improvements to the draft versions. We also wish to thank the NEN Netherlands AI Medical Device Expert Group, The Netherlands Patients Federation and the Health and Youth Care Inspectorate (IGJ) for their input.

Action team members

Jan Jaap Baalbergen (NFU), Robert Geertsma (RIVM, Dennis Japink (ZN), Carl Moons (UMCU), Rozemarijn Pennings (InEen), Marlies Schijven (Amsterdam UMC), Jaap Schrieke (GGZ Nederland), Inge Steinbuch (ActiZ), Jos Schimmelpennink (Nederlandse Vereniging van Ziekenhuizen), Stefan Visscher (Federatie Medisch Specialisten) en Laurine Keulemans (Ministerie van VWS).

Working party members

Phase 1: Paul Agra, Amy Eikelenboom, Christian van Ginkel, Martine de Vries, Saskia Haitjema, Andre Dekker.

Phase 2: Daniel Oberski, Desy Kakiay, Kicky van Leeuwen, Joran Lokkerbol, Evangelos Kanoulas, Gabrielle Davelaar.

Phase 3: Wouter Veldhuis, Bart-Jan Verhoeff, Vincent Stirler, Daan van den Donk, Huib Burger.

Phase 4: Giovanni Cina, Martijn van der Meulen, Maurits Kaptein, Floor van Leeuwen, Egge van der Poel, Marcel Hilgersom.

Phase 5: Teus Kappen, Sade Faneyte, René Verhaart, Jonas Teuwen, Ewout Steyerberg, Leo Hovestadt, René Drost.

Phase 6: Anne de Hond, Bart Geerts, Nynke Breimer, Karen Wiegant, Laure Wynants, Lysette Meuleman.

Reviewers (in alphabetical order)

Annemarie van 't Veen, Charlotte Brouwer, Daniel Vijlbrief, Elise Quik, Jan-Jaap Visser, Jan-Kees van Wijnen, Jan-Willem Wasmann, Jean-Paul Kleijnen, Joris van Dijk, Leon Doorn, Lieke Poot, Maaike van Mourik, Mark Scheper, Martin van Buuren, Merel Huisman, Richard Bartels, Rimmert Brandsma, Rob Tolboom, Roel Streefkerk, Roel van Est, Wouter Bulten.

Participants in field test

RetCAD, Thirona - Mark van Grinsven

HUME, Mentech – Reon Smits and Erwin Meinders

U-Prevent, Ortec – John Jacobs

Predictive model aggression in GGZ (mental health care), UMC Utrecht – Karin Hagoort

Covid-19 severity score, Maasstad Hospital – Sade Faneyte

Board of editors

Maarten van Smeden, Carl Moons, Lotty Hooft, Ilse Kant, Hine van Os, Niels Chavannes, supported by Ylja Remmits en Alexander Boer (KPMG).

Introduction

Authors

Maarten van Smeden, Ilse Kant, Alexander Boer

On behalf of the working party members on Medical AI

Status of this document

This guideline provides a description of what the work field considers good professional conduct in the development, testing and implementation of an *Artificial Intelligence Prediction Algorithm* (AIPA) in the medical sector, including public healthcare. It is up to the work field to determine to what extent the guideline is mandatory in nature. Therefore, compliance with the guideline is not legally binding. The ambition is for this guideline to be accepted as a widely supported guideline. Therefore, the rest of this document will refer to the guideline.

When the need arises, the guideline will discuss obligations arising from applicable legislation and regulations, for example the *General Data Protection Regulation* (GDPR), or requirements for *medical devices* and in-vitro diagnostics as listed in the *Medical Device Regulation* (MDR) or *In-Vitro Diagnostic Medical Device Regulation* (IVDR). However, the guideline does not attempt to offer an exhaustive list of the applicable legislation and regulations, or the guidelines and ISO standards that are customary in medical information technology or the medical devices sector. The guideline thus serves as a supplement to the existing laws and regulations, guidelines, and standards based on what is considered good professional conduct in the field. Furthermore, the guideline does not serve as an elaboration of the pending proposal by the European Commission for an Artificial intelligence Act and the documentation requirements that will be associated with it (in Appendix IV of that proposal) or the legislative proposal on Electronic data exchange in Healthcare (wetsvoorstel Elektronische gegevensuitwisseling in de Zorg (Wegiz)), which will create the opportunity to oblige organisations to exchange certain data in digital format. Both proposals are deemed extremely relevant to the future of the work field. In summary, this guideline should be seen as a supplement to existing laws and regulations, guidelines, and standards.

The framework for supervision on the development, testing and implementation of the AIPA in general falls beyond the scope of the guideline. The guideline is not a prescription, it is not an assessment tool and it is not a risk analysis tool. The guideline first and foremost discusses good professional conduct and not how an organisation can define and monitor good professional conduct in the various environments in which the guideline can be implemented. With regard to the ultimate decision for the type of monitoring process and the documentation, it is very important to determine whether the AIPA will – for example – form part of a medical device as referred to in the MDR and whether it will be marketed in the European Union (Art. 5 MDR), or whether it will be produced and used within a single care institution (the exception in Art. 5 Section 5 MDR). It is also very important to consider which risk class of the MDR the device will be assigned to (according to Rule 11 of Annex VIII of the MDR). The guideline leaves this question open. The guideline contains recommendations that will effectively be mandatory in order to obtain market approval in accordance with the

MDR. In other cases, the guideline expands the existing standards to encompass a broader scope. In order to increase the functionality of the guideline, this document will regularly refer to this existing legislation and regulations where applicable.

Scope

This guideline applies to the development, testing and implementation of an AIPA that forms part of a device intended for use in the healthcare sector, which includes home nursing and self-care. Devices intended for use in the but are not limited to, medical devices as referred to in the MDR. The device can involve independent software, or a device that contains software. For a definition of AI, please refer to Art. 3(1) of the proposal by the European Commission for an Artificial intelligence Act, or the more detailed, but substantively comparable, definition of 18 December 2018 by the *AI High Level Expert Group on Artificial Intelligence* of the European Commission in Box 1.

Box 1: Definition AI by the AI HLEG on 18 December 2018

"Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.

As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization) and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)."

In this guideline, the term AIPA is defined as:

An algorithm that leads to a prediction of a health outcome in individuals. This includes, but is not limited to, the prediction of the probability or classification of having (diagnostic) or developing over time (prognostic) desirable or undesirable health outcomes.

Diagnostic AIPAs predict the probability for individuals in the general population (screening) of – for example – having a certain condition or disease with certain symptoms or complaints, or the probability of having an underlying condition without experiencing any symptoms or complaints.

Prognostic AIPAs predict the probability of occurrence over time of health outcomes in – for example – patients with a certain condition or disease, or they predict the probability of the need to undergo certain treatments or hospitalisations, or they predict for individuals in the general population whether they will develop a certain condition or a certain quality of life over time.

The health outcomes predicted by the AIPA are health outcomes for the *individual* patient, client, or citizen, but can also include health outcomes for third parties, for example the degree of suffering experienced by family members or carers in mental health care. The use of AI other than for the prediction of *individual* health outcomes falls beyond the scope of this guideline, for example navigation applications in robotics, applications that predict patient flows at the population level for use in capacity planning, or applications used solely for predictive classification and segmentation without serving a direct diagnostic or prognostic purpose. However, in such cases, an AIPA can form part of such applications. In these situations, only the AIPA falls within the scope of the guideline. Software *intended to provide information* used in making decisions based on an AIPA, for example for diagnostic, prognostic, therapeutic or prophylactic purposes – including lifestyle adjustment – as referred to in the previously listed Rule 11, generally falls within the scope of the guideline.

The guideline sometimes refers to a *medical context*, which should be interpreted as any conceivable context or interaction in the healthcare sector, regardless of whether a care provider or care institution is involved in this process. In this guideline, the term “medical context” refers to the implementation of an AIPA in Cure, Care and Prevention, including self-service healthcare solutions. Any explicit mention of “medical intervention” refers to procedures restricted to medical professionals as referred to in the *Dutch Individual Healthcare Professions Act*.

Parties involved

The requirements and recommendations in the guideline are addressed directly to the developers and testers of the AIPA, the manufacturer of the software that incorporates the AIPA and the care organisation that implements this software in its organisation. The intended audience includes manufacturers of devices that include an AIPA, researchers developing and testing an AIPA, care organisations and care providers who wish to purchase and use such devices and authorities that help to determine the quality, deployability and reimbursement of the AIPA. The guideline describes what care providers, citizens, patients, patient representatives, insurance companies and policy-makers (such as the National Health Care Institute and the Dutch Healthcare Authority) can expect from an AIPA

developer or manufacturer when they purchase such medical devices, use these devices or have such devices used on them.

The AIPA *developer* or *tester* is any person involved in the development and testing of the AIPA – either in a professional capacity or as a volunteer – and who strives to maintain good professional conduct, for example researchers, data managers, data suppliers, developers and data scientists.

The *manufacturer* is any natural or legal person who manufactures or fully refurbishes a device or has a device designed, manufactured, or fully refurbished, and markets that device under its name or trademark, as referred to in the MDR. The presence of a manufacturer as defined in the MDR does not determine the applicability of this guideline, but the obligations for the manufacturer can apply to the developer in this case.

The *care organisation* is any legal person who makes the device that contains the AIPA available to the end-users and has obligations towards the end-users. The developer or tester of the device can perform his work on behalf of the care organisation that will be using the AIPA. In that case, there is no manufacturer and the roles of manufacturer and care organisation are combined for the interpretation of the requirements and recommendations. The care organisation is usually an organisation in which care providers deliver care – for example a hospital, nursing home, mental health facility or a primary care facility offering accommodation and treatment – but it can also be a welfare organisation or a municipality. The exact interpretation of the definition has deliberately been left open.

The *care provider* is the natural person who provides care – in a professional capacity or as a volunteer – and applies or uses the device that contains the AIPA as end-user in the process. The care provider can act as a care organisation if the care provider does not supply care on behalf of an organisation, for example an individual GP, dentist or psychologist. In that case, there is no care organisation acting as intermediary.

The *patient, client* or *citizen* is the person that is the subject of the prediction by the AIPA. This person can also be the end-user of the device that contains the AIPA. In that case, we generally refer to this as self-care, for example in the (primary) prevention setting.

The *stakeholders* are all parties and individuals who are involved in the development, validation or use of the AIPA, or who are otherwise involved. This includes all aforementioned categories, such as developers and users (e.g. care professionals, patients, citizens) and also auditing and supervisory or certifying parties (e.g. privacy experts, notified bodies, MECs) and the ultimate target groups in or for who the predictions are made (e.g. patients and citizens, depending on the target groups of the AIPA).

Comply or explain

The guideline distinguishes between requirements and recommendations for good professional conduct. The requirements are indicated by **mandatory**. Recommendations are indicated by **recommended** or **strongly recommended**. Use of the guideline implies a *comply or explain* approach, in which the decision whether or not to implement the recommendations is based on a risk assessment that can only be made in line with a specific application of the AIPA. This risk assessment is made explicit and can be explained to third parties. The guideline can be observed if a good explanation is provided, without following all the recommendations. In some cases, the recommendation will clearly state which risks or circumstances would result in the decision to provide an explanation.

The impact of the prediction on the patient, client or citizen is an important consideration when estimating the risks of use, implementation and reimbursement of AIPAs. Good professional conduct remains important even if the expected impact of the AIPA on the patient, client or citizen is low and those sections of the guideline not related to the rights and obligations of final users continue to serve as a guideline for good professional conduct during implementation. This is the case at least if the AIPA will not form part of a device.

Phases of development, testing and implementation

The guideline is divided into six phases:

- Phase 1: Collection and management of the data
- Phase 2: Development of the AIPA
- Phase 3: Validation of the AIPA
- Phase 4: Development of the required software
- Phase 5: Impact assessment of the AIPA in combination with the software
- Phase 6: Implementation and use of the AIPA with software in daily practice.

The proposed chronology is not meant to be imperative and does not always match the factual or most efficient order of actions. For example, the development of necessary software (Phase 4) can already take place at the time of the development of the AIPA (Phase 2). The guideline does serve as a guide for internal supervision, in which the phases serve as a structure to organise the documentation, to determine which AIPA evidence is available/known at a particular time, or to determine which data or evidence is still missing and needs to be collected.

An AIPA solely intended for medical research generally ends after phase 3. The guideline also offers good guidance for professional conduct in this case. Phases 4 through 6 apply more specifically to the manufacturer of the device, the care organisation that will implement the AIPA and the end-users and stakeholders such as the care providers, patients and citizens.

Phase 0: Preparations for the development process

A development process aimed at implementation in health care delivery or self-care generally does not start with the preparation and management of the data in phase 1, but will start with the preparation of the decision to develop the AIPA and the allocation of resources for this. Therefore, it is important in this context that *phase 0: Preparations for the development process* is also mentioned.

During phase 0, the domain experts and end-users jointly determine whether it is necessary to develop an AIPA for the envisaged problem and the feasibility of an idea for development of the AIPA is tested. Generally, these considerations will be based on experiments or a *proof-of-concept (PoC).* An initial information risk assessment will also be performed and a plan of action will be selected in a multi-disciplinary setting, including the required risk mitigation measures and internal supervision. Lastly, an estimate of the total costs and benefits of implementation of the plan can then be made.

In order to make a sound assessment of the most important considerations in phase 0, it is important to gain insight into which parties and individuals have an interest in the AIPA that will be developed. In addition to users, it is advisable to include patients, clients or citizens in the development even in this early phase.

Phase 0 does not form part of the guideline. However, it is important to start thinking in this phase about the implementation of specific standards and recommendations from the guideline. For example, risks and ethical considerations determine which recommendations from this guideline will be implemented and practical considerations will subsequently determine whether the implementation of these recommendations will result in a feasible plan from a business perspective. Therefore, it is important to consider the *comply or explain* choices in the context of the guideline from this phase onwards, as these already start having an effect during this early phase.

No expert sessions in the form of working parties have been organised for phase 0 and no requirements or recommendations have thus been drafted for this phase. Phase 0 therefore falls outside the scope of this guideline.

Creation of the guideline

Preparatory systematic literature review summarised in the article *Guidance and Quality Criteria for Artificial Intelligence based Prediction Algorithms in healthcare: a scoping review*[2]; commissioned by the Ministry of Health, Welfare and Sport, formed the starting point for this guideline.

This extensive literature review was then used to appoint a multi-disciplinary working party for each of the aforementioned six phases, including many experts from the field, such as care providers, standards experts, epidemiologists, data managers, ethicists, statisticians, policy officers, quality officers, data scientists and AI experts employed in fields such as government, teaching and non-teaching hospitals and industry. Within each phase, the working parties placed various relevant topics on the agenda for discussion, prioritisation and further elaboration, to arrive at a guideline of minimum requirements and recommendations. Stakeholders in the field then provided comments on the guideline after being recruited via broad public announcements by the Ministry of Health, Welfare and Sport. A field test was also performed using five different AIPAs and the Patient Federation of the Netherlands, the Netherlands AI Medical Device expert Group of the NEN and the Health and Youth Care Inspectorate provided input.

References

1 A definition of Artificial Intelligence: main capabilities and scientific disciplines | Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines (24 June 2021)

2 Anne A.H. de Hond*, Artuur M. Leeuwenberg*, Lotty Hooft, Ilse M.J. Kant, Steven W.J. Nijman, Hendrikus J. A. van Os, Jiska J. Aardoom, Thomas P.A. Debray, Ewoud Schuit, Maarten van Smeden, Johannes B. Reitsma, Ewout W. Steyerberg, Niels H. Chavannes‡ and Karel G.M. Moons‡. *Shared first author, ‡ shared last author. Guidance and Quality Criteria for artificial intelligence-based prediction models in healthcare: a scoping review. Under review.

1 Collection and management of the data

Authors

Maarten van Smeden, Ilse Kant, Lotty Hooft, Paul Algra, Pieter Boone, Andre Dekker, Amy Eijkelenboom, Christian van Ginkel, Saskia Haitjema, Martine de Vries, Hine van Os, Niels Chavannes, Carl Moons

Phase 1 encompasses the collection and management of the required data for phases 2 through 6. Data can be made available for the development of the AIPA (phase 2), the external validation of the AIPA (phase 3), the software development (phase 4), the impact assessment of the AIPA in combination with the software (phase 5) and the implementation in daily use (phase 6). Phase 1 thus plays an overarching role in the entire process towards implementation and use of the AIPA (with software) in daily practice. The specific requirements for the data can vary per phase. This guideline does not discuss how to decide which data should specifically be collected or used for the development, validation and implementation of an AIPA.

The interpretation of phase 1 centres around drafting, managing and implementing a so-called data management plan. This plan records arrangements, (processing) agreements and procedures for the collection of the required (meta)data, the storage of this (meta)data and the accessibility of this data.

One AIPA, several data management plans

The various AIPA development, testing and implementation phases require different types of data to be collected and/or managed. Different parties are often involved in the data collection for the various phases. An AIPA that has completed all the phases will probably have several data management plans, because each phase has its own (legal and ethical) requirements and its own research methods. When drafting a data management plan, it is wise to consider the expectations of the regulatory and certifying authorities that will require the submission of a data management plan (for example, notified bodies and ethics committees) as part of the procedure for marketing authorisation for medical devices.

The exact design of the data management plan depends on many different factors. In general, we can distinguish four core domains of the data management plan: legal prerequisites, data collection, metadata and availability of data.

In each phase where data will be collected, it is **mandatory** that the developer first drafts a (new version of a) data management plan. **(1a)**

Collection is defined as the gathering or pooling of data to form a data set for use in the development or evaluation of the AIPA, even if this concerns a collection of pre-existing registers or internal data sources. Even the information required by the AIPA in phases 5 and 6, which is provided by the end-user, is ultimately categorised as data collection under the data management plan.

1.1 Legal prerequisites

It is **mandatory** that the developers record the legal prerequisites and context in the data management plan, either by description or by reference. **(1.1a)**

It is **mandatory** that – at the very least – the national and European legislation and regulations that apply to the data and the AIPA based on that data are described. **(1.1b)**

Examples in this context include the *Medical Device Regulation* (MDR), the *Act on the medical treatment agreement* (WGBO), the *Act implementing the NIS directive (WBNI)*, the *Act on medical research involving human subjects* (WMO) and the *General Data Protection Regulation* (GDPR). Other legal prerequisites depend on the intended purpose of the AIPA and the form in which it will be implemented.

In addition, in the case of a collaboration between organisations or the use of data from third parties, it is **mandatory** to record which contracts or agreements exist between the parties (e.g. data processing agreement with external parties), which arrangements have been included in these agreements (e.g. regarding information security and storage period) and any agreements with regard to intellectual property. **(1.1c)**

In addition, it is **strongly recommended** that the existence and operating effect of general information security measures regarding access to data for legal compliance are noted by referring to the appropriate documentation, e.g. an ISO 27001 or NEN 7510 certification. **(1.1d)**

1.2 Data collection

It is **mandatory** that the characteristics of the data collection are recorded in an accurate and detailed manner by the developer in the data management plan that pertains to the specific AIPA development, evaluation or implementation phase. **(1.2a)**

For data collection, it is **mandatory** that at least the following are recorded:

(i) The origin of the data, such as the (expected) start and (expected) end date of the data collection, location(s) of collection (e.g. whether data was collected from hospitals or registers),

(ii) The original objective and the context of the data collection, including the applicable inclusion and exclusion criteria for the target group (e.g. patients or citizens) and in cases where data processing relies on explicit consent from patient, client or citizen, the conditions under which the patient or citizen has given consent (including the processing objective),

(iii) The procedures for measurements and registration of data, such as the design of the data collection (e.g. cohort study, routinely collected care data), the timing of measurements with which data will be collected from individuals (e.g.

measurement of patients immediately after hospitalisation, periodic repetition of measurements) and – if applicable – the technical characteristics of measurement instruments (e.g. manufacturer, model number and sensitivity/responsiveness). **(1.2b)**

The starting point of phase 1 (and the accompanying data management plans) is that these descriptions are sufficiently detailed that the data collection and/or data extraction could – in theory – be reproduced, whether by the developers themselves or by a third party.

1.2.1 Privacy and traceability

As far as privacy is concerned, the current legislation (the current GDPR) will take precedence, irrespective of whether the data pertains to residents of the European Union.

It is **mandatory** that the privacy of persons whose data has been obtained is respected and guaranteed by the developer. **(1.2.1a)**

It is **mandatory** that traceability of data to individuals is prevented (anonymisation) or reduced (pseudonymisation). (**1.2.1b)**

In addition, it is **mandatory** that the principle of data minimisation is followed, meaning that no more data should be recorded per subject than strictly necessary for the development or use of the AIPA. **(1.2.1c)**

In such cases, metadata can be used in combination with data to identify individuals. **It is recommended** that the possibility of re-identification of individuals by means of combining the data about the individual and the metadata about the data is examined and that the results of this research is included in the decision about recording metadata. **(1.2.1d)**

In addition, it is **mandatory** that the AIPA developer or tester – if applicable – explicitly states in the data management plan how they will handle any incidental findings (findings that are uncovered during research that serves another purpose) and the right to destruction of data from individuals whose data have been obtained. **(1.2.1e)**

Data collection in various phases may be subject to the WMO. It is mandatory that research subject to the WMO is always first submitted to a recognised Medical Research Ethics Committee (MEC) or the Central Committee on Research Involving Human Subjects (CCMO) for testing, irrespective of the type of research. In addition, a data protection impact assessment (DPIA) is often required in the context of the GDPR.

It is **strongly recommended** that the plans regarding privacy and traceability are evaluated by conducting a data protection impact assessment (DPIA) or by approaching a privacy expert or MEC, even if no legal obligation to do so exists. **(1.2.1e)**

1.3 Metadata

It is **mandatory** to provide a detailed description of metadata in the data management plan. **(1.3a)**

Metadata is data that provides insight into the characteristics of the collected data – data about data – and describes aspects such as the collection, reporting and accessibility of the collected data. In essence, this involves the recording in general terms of the described characteristics and processes (as described in 1.2) in the data collection process itself. This should focus on providing transparency and clarity about the collected data.

It is **strongly recommended** that metadata is recorded at the following levels:

- *Data provenance*[1] (i.e. data lineage): contains information about the origins of the collected data (points), any changes and transformations to the data, including classification of the purpose of the change and other details that can provide information about the validity of the collected data, insofar as this is compatible with the recommendations regarding traceability in 1.2.1.
- *Medical context:* information about the design of the data collection and the population context (e.g. consecutive patients visiting the GP with skin complaints, hospital patients referred for a CT due to suspected pulmonary embolism, healthy individuals in the general population aged 70 years and older to determine how high their risk is of developing a certain type of cancer). In addition, this also describes the physical and social environmental determinants of the included population, which were relevant to the implementation of the AIPA.
- *Characteristics and descriptive statistics of the data,* such as the units, averages, ranges of values, description of missing observations and any shifts or trends over time. **(1.3b)**

It is **mandatory** to base the choice of metadata and the description of metadata on an inventory of the interests of the various stakeholders who should be granted access to the metadata, in particular the inspection authority or certifying bodies and – in the event of collaborations – partner (care) organisations. **(1.3d)**

If multiple data sources will be used in a certain phase, for example different data sets from different data collection processes for the validation (phase 3) of the AIPA, then it is **strongly recommended** that the metadata is presented separately for each data source, specifying how the sources are linked. **(1.3e)**

1.4 Availability of data

It is **mandatory** to provide clear information about the availability of the data, for stakeholders and third parties in the data management plan. **(1.4a)**

It is **strongly recommended** that the FAIR principles[2] are followed when making data available (internally or externally). **(1.4b)**

FAIR is an acronym for *Findable*, *Accessible*, *Interoperable* and *Reusable*. The FAIR principles are guidelines for the description, storage and publication of (meta)data.

In the event that data is made available to partners or third parties, it is **mandatory** to record the agreements about the storage of the used data in the data management plan. This should include at least: the form in which the data will be stored, the location(s) for data storage, the scheduling of incidental and periodic data back-ups, agreements about potential incidents such as data leaks and the (remaining) storage period for the data. **(1.4c)**

Where applicable, it is **mandatory** to record the process of ensuring compliance with the current national and international legislation and regulations regarding the processing of personal data, data storage and data security, as described in – among others – the General Data Protection Regulation (GDPR) and the Network and information systems security act (WBNI) and pursuant guidelines. **(1.4d)**

In addition, it is **recommended** that the data is made available in forms that are in line with information standards that are commonly used in digital information exchange in the health care sector. **(1.4e)**

Examples for the Netherlands include the various information standards for information exchange in health care that are managed in the Dutch care setting by Nictiz[3] and play an important role in the exchange of patient data between various care institutions and care providers. An example of an international standard is the medical terminology system SNOMED[4].

In the event that data made available contains information that can be traced to individuals and for interpretation of the aforementioned requirements, the data management plan can refer to the data processing agreements that have been concluded, insofar as these agreements list the arrangements in place for all these points.

1.5 Version management and availability of the data management plan

It is **mandatory** for the data management plan to be made available by the developer to the parties involved in the data collection or data processing. **(1.5a)**

It is **recommended** that the data management plan is made publicly accessible or accessible on request, for example by publishing it on a publicly accessible website. **(1.5b)**

This recommendation may be weighed against the commercial interests.

It is **mandatory** for version management to be implemented for all components of the data management plan. **(1.5c)**

This means that any changes to the data management plan must be accurately logged and recorded over time. This means that the data management plan is a living document, which will be regularly updated in the subsequent phases or drafted from scratch in a subsequent phase.

1.6 References

1. Gupta A. Data Provenance. In: Liu L, Özsu MT, eds. *Encyclopedia of Database Systems* Boston, MA: Springer US; 2009. P. 608–608
2. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, Silva Santos LB da, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, Hoen PAC 't, Hooft R, Kuhn T, Kok R, Kok J, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018.
3. Nictiz. Standaardisatie van digitale gegevensuitwisseling in de zorg. https://www.nictiz.nl/overig/standaardisatie-van-digitale-gegevensuitwisseling-in-de-zorg/. Published December 5, 2018. Accessed December 8, 2021.
4. SNOMED. SNOMED International. https://www.snomed.org/. 2021. Accessed December 8, 2021.

2 Development of the AIPA

Authors

Maarten van Smeden, Ilse Kant, Lotty Hooft, Gabrielle Davelaar, Desy Kakiay, Evangelos Kanoulas, Kicky van Leeuwen, Joran Lokkerbol, Daniel Oberski, Hine van Os, Niels Chavannes, Carl Moons

Phase 2 covers the development of the AIPA model. The model is the entire set of algorithm-specific data structures that forms the AIPA in combination with an algorithm and is the result of analysis of the training data. This document does not provide a specific step-by-step plan for the analytical AIPA model development; the reader should refer to existing literature for this.[1–6].

The use of a standardised step-by-step plan is **strongly recommended** for the complete recording of the development steps and the procedures and results of internal validation (see below) of the AIPA. **(2a)**

The TRIPOD reporting guidelines[7–9] (www.tripod-statement.org) serve as a guide in this process and a specific TRIPOD-AI reporting guideline is nearing completion.

2.1 Explanation of intended use

It is **mandatory** for the developer of the model to define and record a clear definition of the intended use of the AIPA. **(2.1a)**

It is **mandatory** to clarify at least the following in the intended that has been recorded:

i) For which medical or health application the AIPA is intended (e.g. in which medical context, indication or target population) and who the envisaged end-user is (e.g. a specific specialisation, primary care provider, or the patient, client or citizen himself);

ii) Which medical or healthcare process the AIPA intends to influence and what the expected benefit is compared to the current process (e.g. faster diagnosis, improved estimate of a person's prognosis, or indication for modification of a lifestyle habit);

iii) What the envisaged timing of the use of the AIPA or the prediction will be (e.g. upon admission to the hospital or Intensive Care Unit, at the time of receiving a cancer diagnosis, upon referral for a CT scan, or when symptoms or complaints are observed, or when monitoring blood sugar levels);

iv) Whether this is a diagnostic, prognostic, monitoring, screening or other type of healthcare application;

v) What the prediction horizon of the AIPA is (in the case of prognostic predictions: how far forward in time does the AIPA prediction go). **(2.1b)**

Involvement of stakeholders – such as users and patients, clients or citizens – in defining the intended use is **strongly recommended**. **(2.1c)**

1.1.1. Data set(s) and intended use

An exact description of the origin of the data set(s) (e.g. time/place) used in the development of the AIPA model, the design of data collection (e.g. consecutive patients), measurement and registration procedures, any selections, inclusion and exclusion criteria of participants or data points in the analyses has already been provided in phase 1.

In general terms, the use of a representative sample from the target population (as recorded in the intended use, refer to section 2.1) for the development of the AIPA is **strongly recommended**. **(2.1.1a)**

If the data used is not (completely) representative, or this is suspected, then it is **mandatory** for this to be documented and substantiated. **(2.1.1b)**

2.2 Analysis and modelling steps

It is **mandatory** for the developer of the model to record all analysis and model development steps. This includes all preparatory steps (e.g. initial data analysis[10], feature engineering), modelling technique used (e.g. neural network, random forest, time to event, logistic regression), all modelling steps (e.g. model selection, tuning, (re)calibration). **(2.2a)** The starting principle is that the consecutive analysis and modelling steps are sufficiently detailed to ensure that a third party would be able to reproduce the data exactly based on the description of all the analysis and modelling steps [7–9, 11].

2.3 Internal evaluation of the model

2.3.1 Internal validation

Internal validation is an important part of the process of development of the AIPA. The aim of the internal validation is to quantify realistic estimates of the model performance of the AIPA. An adequate estimator of the model performance (e.g. the C(oncordance) statistic and calibration curve[8,12]) can differ between types of applications and endpoints (e.g. binary, multi-category, *time-to-event*), also refer to section 3.1.2. Explicit minimum criteria for model performance have not been provided in this document, because minimum model performance is dependent on context.

Description of the model performance in context is **strongly recommended**, e.g. by comparison to other predictive models or AIPAs for the same medical context or target population, or by comparison to a benchmark relevant to the medical context, so that it becomes possible to assess the benefit compared to the current medical practice. **(2.3.1a)**

It is **mandatory** to implement adequate measures to minimise *optimism about model performance*[1,2,5,6] in order to achieve realistic estimates of model performance. **(2.3.1b)**

This means that the internal validation must be strictly separated from the model development, for example the variable and model selection and tuning of the model (i.e. be wary of *leakage*). For example by performing *nested cross-validation*, in which the execution of all model development steps (*inner loop*) is separated from the internal validation of the model (*outer loop*).

The use of statistically efficient internal validation methods (e.g. cross-validation, bootstrap) – in which all available data are used for the development of the model – is **strongly recommended** instead of the use of inefficient internal validation methods (e.g. several train-test splits)[13]. **(2.3.1c)**

Any deviation from this recommendation, for example due to lack of computational feasibility, must be substantiated.

2.3.2 Analysis of potential (negative) impact of the model

In addition to realistic estimates of model performance, it is important to look ahead throughout the development process at the (potential) application of the AIPA in practice, so that the development of the AIPA continues to match the medical context and the problem that needs solving.

Performance of a credible and transparent analysis of the potential negative impact of the use or implementation of the AIPA is **recommended** and this should be recorded as part of the assessment of the benefit of the AIPA compared to current medical practice. **(2.3.2a)**

For example by performing an analysis of the predictive errors of the model (i.e. *error analysis*) and relating this explicitly to the intended use.

It is **recommended** that an estimate of *fairness* risks should be performed together with stakeholders from the medical context envisaged for the intended use. Refer to Section 3.3 for a detailed elaboration. **(2.3.2b)**

Model performance is not always equal for all generalised sub-populations.

Therefore, it is **strongly recommended** that the *heterogeneity* in estimated model performance of the AIPA should be determined in as much detail as possible, for example by use of data from several locations (e.g. various medical centres) or other patient-relevant context – for example using *internal-external cross-validation*[14,15]. **(2.3.2c).**

Any deviation from this should be substantiated with reference to the intended purpose of the model and an assessment of the risks to the robustness of the model.

Examination and recording of the expected benefit in the medical context of the model is also **strongly recommended**. **(2.3.2d)**

This can be achieved, for example, using *a decision curve analysis*[16]. Another – more robust and comprehensive – method to study the impact on the medical practice at an early stage of the development of an AIPA is by an *early Health Technology Assessment (eHTA)* of the AIPA[17,18].

2.4 Technical Robustness

During the development of the AIPA, it is **mandatory** to examine the technical robustness of the model and record the findings in a transparent manner, at least for those models that are used in the external validation (phase 3). **(2.4a)**

It is **strongly recommended** to use technical robustness as a criterion for model selection, in addition to model performance (as referred to in 2.3.1). **(2.4b)**

Various sensitivity analyses are recommended in order to study the robustness. This can include analyses of the:

- *Architectural robustness:* repetition of the analysis steps on the same data results in a model that does not deviate significantly from the original model.
- *Consistency of model performance:* repetition of the analysis steps on the same data results in models with performance that does not deviate much from the performance of the original AIPA.
- *Adversarial robustness:* the effect of a (deliberate) disruption on the input variables of the model on the performance and/or architecture.
- *Domain shift and outliers:* the effect of any *outliers* in the data and/or deliberate changes to the data set (e.g. deliberate inclusion or exclusion of certain groups) on the model performance and/or the architecture (e.g. *outlier rejection* analysis). Also refer to phase 4 and 6 for additional activities.

In addition – to increase the transparency of the AIPA – insight can be provided on the effect of certain input variables on the performance, for example, by using *feature importance* methods (i.e. *explainable* AI[19]).

In addition to the analyses of the technical robustness of the AIPA model during the development, the robustness of the model in combination with the software that the model forms part of should also be analysed. Refer to phase 4 for this.

2.5 Size of the data set for development of the AIPA

The guiding principle for selecting the size of the data set for development of the model is: the bigger, the better. However, this guiding principle should be weighed against medical ethics considerations and the requirement of data minimisation from phase 1. In general, the minimal required size of the data set increases as the incidence (or prevalence) of the outcome requiring prediction lies further away from 50% (i.e. higher *class imbalance*), as there are fewer strong predictors for the outcome in the input (lower explained variance in the outcome as a result of the input variables) and as the model contains more input variables and/or becomes computationally more complex. For models based on regression, there are explicit rules and formulas that can be applied to calculate the minimum size of the data set[20,21]. These types of rules for *a priori* calculations of minimum size of the data set do not exist (yet) for more complex models. However, *a posteriori* sample size criteria[22] – e.g. so-called *learning curves*[23] – do exist, which can be used to evaluate whether the data set meets minimum criteria, such as a limited risk of over-fitting and exact estimate of the individualised outcome probabilities.

The use of *a priori* or *a posteriori* methods to evaluate whether the size of the data set meets the minimum criteria is **strongly recommended**. **(2.5a)**

An MEC will ask for a justification of the size of the data set.

2.6 Logging, availability and version management

2.6.1 Logging, reproducibility and replicability

Reproducibility and replicability are important guiding principles for the development of the AIPA.

It is **mandatory** for all analysis steps (also refer to Requirement 2.2a) and internal validation steps and the analysis of technical robustness to be logged accurately in order to guarantee the reproducibility (i.e. the ability to repeat the development using different data). **(2.6.1a)**

The guiding principle is once again that the logging should be sufficiently detailed to allow third parties to reproduce the development steps. The TRIPOD reporting guidelines[7-9] (www.tripod-statement.org) can be used as a guide in this process.

In addition, where applicable, it is **recommended** that results are published in a scientific journal. **(2.6.1b)**

It is also **recommended** that computer codes used in the AIPA development are published or made available on request, so that these can be used by third parties for an independent validation of the AIPA model[24]. **(2.6.1c)**

In addition, it is **recommended** that the replicability by third parties (i.e. repeating the model development using the same data) is guaranteed by making the data available on request if possible. **(2.6.1d)**

Of course due consideration must be given to the current legislation and regulations regarding privacy and the pursuant limitations and risks, such as the risk of identification of those involved. These recommendations may also be weighed against commercial interests and it is possible to argue that the external validation should be performed by a trusted third party.

2.6.2 Version management and availability of the model

It is **mandatory** that the version history (the final model and any updates to the model) is logged accurately, for example by assigning a version number. **(2.6.2a)**

This version history of the model serves as an addition to the version history of the software, as required – for example – by the MDR.

In addition, making the actual model and/or models (e.g. modelling coefficients if applicable, nomograms, computer code (with the actual model)) publicly accessible is **strongly recommended**, including – if available – the (minimum) software surrounding the model for demonstration purposes. **(2.6.2b)**

Any deviation from this recommendation must be substantiated. Commercial interests can form a deciding factor in this deliberation.

2.7 References

1. Harrell FE. Regression Modeling Strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.
2. Steyerberg EW. Clinical Prediction Models. Cham: Springer International Publishing; 2019.
3. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning The Elements of Statistical LearningData Mining, Inference, and Prediction, Second Edition. Springer Ser. Stat. 2009.
4. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, Massachusetts: The MIT Press; 2016.
5. Riley RD, van der Windt D, Croft P, Moons KGM. Prognosis Research in Health Care: Concepts, Methods, and Impact. Oxford University Press, 2019.
6. Moons KG, Kengne AP, Woordward M, Royston P, Vergouwe Y, Altman DG, Grobbee DE. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* 2012;98:683-690.
7. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med* 2015;162:55.
8. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Ann Intern Med* 2015;162:W1–W73.
9. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet* 2019; 393:1577-1579.
10. Huebner M, Vach W, Cessie S le. A systematic approach to initial data analysis is good research practice. *J Thorac Cardiovasc Surg* 2016;151:25–27.
11. Collins GS, van Smeden M, Riley RD. COVID-19 prediction models should adhere to methodological and reporting standards. European respiratory journal 2020; 56: 2002643.
12. Van Calster B, McLernon DJ, Smeden M van, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019;17:230.

13. Debray TP, Riley RD, Rovers MM, Reitsma JB, Moons KG; Cochrane IPD Meta-analysis Methods group. Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: guidance on their use. *PloS Med.* 2015; 12:e1001886.

14. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal–external, and external validation. *J Clin Epidemiol* 2016;69:245–247.

15. Debray TP, Damen JA, Riley RD, Snell K, Reitsma JB, Hooft L, Collins GS, Moons KG. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. *Stat Methods Med Res.* 2019;28:2768-2786.

16. Vickers AJ, Calster B van, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019;3:18.

17. Jenniskens K, Lagerweij GR, Naaktgeboren CA, Hooft L, Moons KGM, Poldervaart JM, Koffijberg H, Reitsma JB. Decision analytic modeling was useful to assess the impact of a prediction model on health outcomes before a randomized trial. *J Clin Epidemiol* 2019;115:106-115.

18. Van Giessen A, Wilcher B, Peters J, Hyde C, Moons KG, de Wit GA, Koffijberg H. Health economic evaluation of diagnostic and prognostic prediction models. A systematic review. *Value in Health* 2014;17:A560.

19. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion* 2020;58:82–115.

20. van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans MJ, Reitsma JB. Sample size for binary logistic prediction models: beyond events per variable criteria. *Statistical methods in medical research* 2019;28:2455-2474.

21. Riley RD, Ensor J, Snell KI, Harrell FE, Martin GP, Reitsma JB, Moons KG, Collins GS, van Smeden M. Calculating the sample size required for developing a clinical prediction model. *Bmj* 2020;368: m441.

22. Balki I, Amirabadi A, Levman J, Martel AL, Emersic Z, Meden B, Garcia-Pedrero A, Ramirez SC, Kong D, Moody AR, Tyrrell PN. Sample-Size Determination Methodologies for Machine Learning in Medical Imaging Research: A Systematic Review. *Can Assoc Radiol J* 2019;70:344–353.

23. Christodoulou E, Smeden M van, Edlinger M, Timmerman D, Wanitschek M, Steyerberg EW, Van Calster B. Adaptive sample size determination for the development of clinical prediction models. *Diagn Progn Res* 2021;5:6.

24. Community TTW, Arnold B, Bowler L, Gibson S, Herterich P, Higman R, Krystalli A, Morley A, O'Reilly M, Whitaker K. The Turing Way: A Handbook for Reproducible Data Science. Zenodo; 2019.

3 Validation of the AIPA

Authors

Maarten van Smeden, Ilse Kant, Lotty Hooft, Huib Burger, Daan van den Donk, Vincent Stirler, Bart Jan Verhoeff, Wouter Veldhuis, Hine van Os, Niels Chavannes, Carl Moons

Phase 3 covers the (external) validation of the AIPA developed in phase 2. External validation refers to the evaluation of the performance of the AIPA model using data that was not used in the (further) development in phase 2[1–3]. We distinguish between the evaluation of the statistical or predictive value, the evaluation of the (added) value compared to the current care practice and the evaluation of *fairness and algorithmic bias*.

The transition from phase 2 to phase 3 is based on the assumption that the development of the AIPA model has been completed. However, in some cases, a small set of candidate models need to be validated in phase 3 to arrive at a final decision for a model. External validation explicitly does not refer to: the (re-)training or (re-)tuning of a model. External validation an result in complete or partial re-training of a developed AIPA. This is also called *model updating*.[4–6] In that case, it is necessary to complete (a part of) phase 1 and 2 again and to log this.

3.1 Evaluation of predictive (statistical) characteristics of the AIPA

3.1.1 Target population and context

To achieve a good evaluation of the predictive value of an AIPA, it is **mandatory** that the tester uses a different data set for external validation than the set that was used for the AIPA development in phase 2, but this data set must be representative for the intended population and context[3]. **(3.1.1a)**

If a so-called *holdout* data set exists, then the characteristics of this data set have already been fully defined in phase 1. If not, at least the characteristics of the data collection process have been recorded.

The use of study designs in which only data from healthy controls are used that are not representative of the target context or population usually results in an excessively optimistic evaluation of the predictive value of the AIPA model (*spectrum bias*[7]).

It is **strongly recommended** to avoid the use of a study design in which only data from so-called healthy controls are used. **(3.1.1b)**

It is **mandatory** for an exact description of the origin of the data (e.g. time and place), the method of data collection (e.g. consecutive patients), measurement and registration procedures, any selections and inclusion and exclusion criteria to be recorded in the data management plan in order to define the context of the predictive characteristics (see phase 1). **(3.1.1c)**

In addition, the intended use of the AIPA should also be monitored closely, as set out in phase 2.

It is **recommended** that exclusion of individuals who do belong to the target population or context should be avoided. **(3.1.1d)**

Any unforeseen exclusion of data or individuals, for example due to failed measurements or the withdrawal of consent, should be accurately recorded and – preferably – described per case or per group.

For the sake of general applicability, the target population or context for external validation can differ from the intended population as formulated and used in the development of the AIPA model (phase 2).

In the event of structural differences between the development (phase 2) and the external validation (phase 3) in terms of the design of data collection, measurement and registration procedures, any selections and inclusion and exclusion criteria, it is **mandatory** to record the reason for this difference in intended populations between phase 2 and 3. In addition, the nature of the differences must also be clearly recorded in the data management plan.[8,9] **(3.1.1e)**

In addition, it is **recommended** – if possible – to compare the baseline characteristics (e.g. distribution in age, gender, co-morbidity) and perform statistical tests between the data used for the development (phase 2) and for the external validation in phase 3 (even if the target populations are the same)[10]. **(3.1.1f)**

This allows any *data drift* to be quantified in later phases.

3.1.2 Model performance of the AIPA

A realistic evaluation of the model performance, i.e. the correspondence between the predicted outcome and the observed outcome, forms an important part of the external validation of the AIPA model.

During the evaluation of model performance, it is **mandatory** to take into consideration the scale on which the predictions are made when selecting estimators. **(3.1.2a)**

A correct estimator or measure of model performance can differ for a binary endpoint (e.g. health outcome is present versus absent), multi-category endpoint (e.g. definitely present, probably present, probably absent, definitely absent), a *survival* endpoint (with possible *censoring*) or an outcome on a continuous scale.

For the selection of estimators of model performance, it is also **mandatory** to take into consideration the predicted output of the AIPA model. **(3.1.2b)**

For example: for predictive models that only provide binary (yes versus no, or present versus absent) classifications as output, the focus rests on the accuracy of classifications and

associated measures such as the F1-score and *C-statistic*, whilst for a predictive model with probability/risk output, the focus often rests on the calibration and discrimination (*C-statistic*) of the model. Please refer to the TRIPOD guidelines[8,9] for a guide on recording these choices.

As for the internal validation in phase 2, it is **strongly recommended** that the estimates for model performance are placed in the intended context as far as possible, for example by comparison of the predictive value to comparable predictive models used for the same context or target population or a relevant benchmark for the medical setting envisaged in the intended use. **(3.1.2c)**

It is also **strongly recommended** that the estimates for model performance of the AIPA during external validation be compared to the model performance reported after internal validation during development (phase 2). **(3.1.2d)**

A large discrepancy in model performance that is uncovered during external validation in phase 3 can be indicative of *over-fitting* of the model during the development in phase 2[5,6].

3.2 Evaluation of medical characteristics and expectations for implementation of the AIPA

As is the case in phase 2 during the internal validation, it is **mandatory** that an external validation of the AIPA model also evaluates the medical characteristics, or the performance in the intended medical setting. **(3.2a)**

An analysis of anticipated costs and benefits is **recommended**. **(3.2b)**

As was the case in phase 2, this can be achieved – for example – using *a decision curve analysis*[11]. Another – more robust – method to study the impact compared to the current medical practice using an early Health Technology Assessment (eHTA[12,13]), as mentioned previously in phase 2.

It is **recommended** that you make an estimate of expected barriers before implementation of the AIPA and record these in this phase, for use in phase 5 and 6. **(3.2c)**

For example by a description of limitations caused by availability of data, an estimate of the time (e.g. entry of data or the calculation time of the model) and costs (e.g. measurements) required to use the model in practice and expected barriers relating to the implementation of the AIPA in the current processes of the envisaged medical practice.

It is **strongly recommended** that stakeholders from the medical setting who are envisaged for the intended use (both end-users and patients) are involved in the evaluation of medical characteristics, the analysis of costs and benefits and the assessment of barriers. **(3.2d)**

3.3 Fairness and algorithmic bias

The external validation of the AIPA model should look beyond the model performance and medical value. The evaluation of *fairness*[17] and bias is also very important.

Unfair treatment is usually the result of a form of algorithmic bias. We can distinguish various forms of algorithmic bias. The terminology framework by Suresh & Guttag (2020)[14] serves as a guide in this matter. They distinguish between six forms of bias:

- *Historical bias*: undesirable model outcomes or predictions as a result of data from the world as it is or as it was. This can be caused, for example, by the fact that the AIPA model was developed using data in which systematic under-diagnosis or over-diagnosis played a role.
- *Representation bias*: undesirable model outcomes or predictions as a result of under-represented sub-groups in the data. This can be caused, for example, by the fact that the AIPA model was developed using data that was not representative for the target population or context.
- *Measurement bias*: undesirable model outcomes or predictions as a result of the AIPA being trained using data in which the outcome variable contained misclassifications (refer to Section 3.4) or due to differences between the (accuracy of the) measurement of predictors/features for the development of the AIPA and the external validation/implementation. This is also referred to as measurement heterogeneity[15,16].
- *Aggregation bias*: undesirable model outcomes or predictions for certain sub-groups. This can be caused, for example, by the fact that the model performance of the AIPA is much poorer in (often under-represented) sub-groups.
- *Evaluation bias*: distorted statistical evaluation as a result of external validation of the AIPA using a data set that is not representative for the target population. Examples include the use of an AIPA in General Practice that was trained using data from hospitals in which more severe disease occurs.
- *Deployment bias*: mismatch between the problem that the AIPA is trying to solve and the way in which it is used by others.

Fairness of an algorithm is generally studied in a risk-oriented approach: hypotheses are first formulated about groups that could potentially be treated unfairly by the AIPA, for example, by means of a systematic analysis of the “guidance ethics approach” (aanpak begeleidingsethiek)[24] (Structured, collaborative and multidisciplinary approach for applying ethics to the development or implementation of technology) and these hypotheses are then analysed during the external validation. At the very least, this takes into consideration groups that can be distinguished based on sensitive personal data under the GDPR.

It is **mandatory** to analyse and record the presence of certain types of bias (such as *algorithmic bias*) that can result in unfavourable disparities in outcome for certain groups in the population . **(3.3a)**

In addition, it is **mandatory** to analyse and record the risk of unequal treatment or undesirable disparities in outcome for certain groups in the population. **(3.3b)**

It is **strongly recommended** that stakeholders such as end-users and patients become involved in this evaluation of fairness risks, for example with regard to the "guidance ethics approach" (aanpak begeleidingsethiek)[24]. **(3.3c)**

3.4 Determining the outcome variable (labelling)

The accurate determination of the outcome that is to be predicted in the external validation data set is an important factor for the validity of the statistical model performance and the medical value. There are many situations in medicine where no gold standard is available for the measurement of the outcome variable (e.g. for some diagnoses, classifications or cause-specific mortality), which potentially result[18,19] in misclassification of the outcome variable. Therefore, the term "reference standard" is often used. In some situations, evaluation by an expert or group of experts is required to arrive at a decision per case (e.g. the assessment of a tumour on a CT scan[20]).

It is **mandatory** to perform so-called *labelling* of outcomes in the data set for external validation as accurately as possible in this phase and to record and justify this process as transparently as possible. **(3.4a)**

It is **recommended** that the process is recorded as accurately as possible, with a report of which experts were involved in the labelling (e.g. training, expertise), in which circumstances (e.g. number of experts per case, available time) and how any discrepancies between labels were resolved. **(3.4b)**

It is **strongly recommended** that the quality of the labels should be quantified (e.g. by means of *measures of agreement* (e.g. the *kappa statistic* or ICC), or by estimating the accuracy of the labelling[21]). **(3.4c)**

3.5 Size of the data set for external validation

The guiding principle for selecting the size of the data set for external validation is: the bigger, the better. The bigger the data set, the more accurate the estimates that will be used for the statistical and medical evaluation and the better the algorithmic bias can be analysed. However, it is important not to loose sight of the importance of accurate labels (refer to

Section 3.4). Please refer to the literature for a calculation of the minimum size of a data set[22,23].

It is **mandatory** to provide argumentation for the size of the data set for external validation. **(3.5a)**

It is **recommended** to perform calculation of the minimum size of the data set, if possible (refer to Section 2.5). **(3.5b)**

3.6 Logging, reproducibility and replicability

As for phase 2, reproducibility and replicability are important guiding principles for external validation of the AIPA.

In order to guarantee *reproducibility* (i.e. repeat of the external validation using different data), it is **mandatory** to log the process that was followed and the data that was used for external validation in a complete and transparent manner[8,9], even in the case of negative results. **(3.6a)** The TRIPOD reporting guidelines[7-9] (www.tripod-statement.org) serves as a guide in this process.

It is **recommended** that the computer codes used for the external validation are made publicly accessible. **(3.6b)**

It is **recommended** that the data is made (publicly) accessible in order to increase replicability (i.e. repeating the external validation using the same data). **(3.6c)**

Of course due consideration must be given to the regulations regarding privacy and the pursuant limitations. These recommendations may also be weighed against commercial interests and it is possible to argue that the external validation should be performed by a trusted third party.

3.7 References

1. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;**19**:453–473.
2. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;**338**:b605–b605.
3. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, Collins GS. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016;i3140.
4. Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, Woodward M. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012;**98**:691–698.
5. Harrell FE. Regression Modeling Strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.
6. Steyerberg EW. Clinical Prediction Models. Cham: Springer International Publishing; 2019.
7. Usher-Smith JA, Sharp, SJ, Griffin SJ. The spectrum effect in tests for risk prediction, screening, and diagnosis. *BMJ* 2016;i3139.
8. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med* 2015;**162**:55.
9. Moons KGM, Altman DG, Reitsma JB, Ioannidis JP a, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Ann Intern Med* 2015;**162**:W1–W73.
10. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2015;**68**:279–289.
11. Vickers AJ, Calster B van, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019;**3**:18.
12. Jenniskens K, Lagerweij GR, Naaktgeboren CA, Hooft L, Moons KGM, Poldervaart JM, Koffijberg H, Reitsma JB. Decision analytic modeling was useful to assess the impact of a prediction model on health outcomes before a randomized trial. *J Clin Epidemiol* 2019;**115**:106–115.
13. Van Giessen A, Wilcher B, Peters J, Hyde C, Moons KG, Wit GA de, Koffijberg H. Health economic evaluation of diagnostic and prognostic models. A systematic review. *Value in Health* 2014;**17**:A560.

14. Suresh H, Guttag JV. A Framework for Understanding Unintended Consequences of Machine Learning. *ArXiv190110002 Cs Stat* 2020;
15. Luijken K, Groenwold RHH, Van Calster B, Steyerberg EW, Smeden M. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective. *Stat Med* 2019;sim.8183.
16. Luijken K, Wynants L, Smeden M van, Van Calster B, Steyerberg EW, Groenwold RHH, Timmerman D, Bourne T, Ukaegbu C. Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *J Clin Epidemiol* 2020;**119**:7–18.
17. Ethics guidelines for trustworthy AI | Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (25 June 2021)
18. Rutjes AWS, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PMM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol Assess* Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academical Medical Center, University of Amsterdam, The Netherlands; 2007;**11**:ix–51.
19. Naaktgeboren CA, Groot JAH de, Rutjes AWS, Bossuyt PMM, Reitsma JB, Moons KGM. Anticipating missing reference standard data when planning diagnostic accuracy studies. *BMJ* 2016;i402.
20. Bertens LCM, Broekhuizen BDL, Naaktgeboren CA, Rutten FH, Hoes AW, Mourik Y van, Moons KGM, Reitsma JB. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. *PLoS Med* 2013;**10**:e1001531.
21. Jenniskens K, Naaktgeboren CA, Reitsma JB, Hooft L, Moons KGM, Smeden M van. Forcing dichotomous disease classification from reference standards leads to bias in diagnostic accuracy estimates: A simulation study. *J Clin Epidemiol* 2019;**111**:1–10.
22. Riley RD, Debray TPA, Collins GS, Archer L, Ensor J, Smeden M, Snell KIE. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat Med* 2021;sim.9025.
23. Archer L, Snell KIE, Ensor J, Hudda MT, Collins GS, Riley RD. Minimum sample size for external validation of a clinical prediction model with a continuous outcome. *Stat Med* 2021;**40**:133–146.
24. ECP. Handleiding aanpak begeleidingsethiek voor AI in de zorg. https://begeleidingsethiek.nl/publicaties/handleiding-aanpak-begeleidingsethiek-voor-ai-in-de-zorg/. Published January 15, 2021. Accessed December 8, 2021.

4 Development of the required software

Authors

Ilse Kant, Maarten van Smeden, Hine van Os, Egge van der Poel, Floor van Leeuwen, Pieter Boone, Giovanni Cina, Maurits Kaptein, Marcel Hilgersom, Martijn van der Meulen, Lotty Hooft, Carl Moons, Niels Chavannes

The development and validation of the AIPA as a predictive model were discussed in phase 2 and phase 3. Phase 4 covers the further development of the software surrounding the AIPA by the manufacturer. In other words: the design, the development, the user testing and the accompanying system requirements of the software that the AIPA will form part of (hereinafter referred to as: AIPA software) also form part of this phase. These system requirements include utilities that must be provided as part of the design of a quality management system. In addition, the information supplied with the software also belongs to this phase.

It should be noted that the developer and the care organisation where the AIPA will be implemented can be one and the same party. In such cases, there will be no interaction between a manufacturer and a care organisation. In that case, the requirements and recommendations for the manufacturer should be interpreted as requirements and recommendations for the developing care organisation and any apparent duplications resulting from this situation can be ignored.

From this phase onwards, it is assumed that the software which the AIPA forms part of is designated for use in the intended medical context, including self-care. Phase 4 will often take place after the external evaluation (phase 3), but can also take place after phase 2. In such cases, both will take place simultaneously, for example because the evaluation of the AIPA can only take place in combination with the software or the device containing the software, which the AIPA forms part of.

4.1 Explainability, transparency, design and information

4.1.1 Explainability, transparency and design of the AIPA software

The outcomes of the AIPA model will be presented in the software in a transparent and explainable manner. The presentation of the outcomes of the AIPA model in the software distinguishes between an inherently explainable and a complex model.

An inherently explainable model (e.g. a decision tree or algorithm in which the weights of the input variables are clear) is a model that allows for direct interpretation of how the model predictions (or classifications) were established.

In the case of an inherently explainable model, it is **mandatory** that the manufacturer discloses information about the interpretation of the model and the model predictions for the intended end-users. **(4.1.1a)** It is **mandatory** that this takes place in a presentation of the model predictions by the software for the end-user, particularly if medical decisions are made based on the model predictions. **(4.1.1b)**

This can be achieved, for example, by providing an explanation in the user interface, where the prediction or outcome is also presented.

In the case of a complex model (e.g. an algorithm based on *deep learning*), the relationship between input variables and predicted outcomes is so complex that it is no longer possible to comprehend this (so-called "black box" algorithms). Extra attention should be paid to post-hoc information and interpretation of the model in the presentation of the model by the Software[1-3] .

It is **mandatory** to substantiate the following aspects of complex models: 1) why an explainable model was not used and 2) if opting for a post-hoc explanation, why this is appropriate for the model and the intended end-user. **(4.1.1c)**

It is **strongly recommended** that in both cases, an amended model presentation and explanation is designed for each end-user and that training aids in the field of interpretation of the AIPA are developed and made available (phase 6) to avoid incorrect model interpretations by the intended end-user. **(4.1.1d)**

The following should be taken into consideration here: usability and testing of the software that the AIPA model forms part of according to the existing standards and regulations (refer to Section 4.4), presentation of the predicted outcomes including information about (un)certainties (for example by providing a confidence interval), use of language appropriate for the end-user (for example, this would differ for a care provider and a patient), intuitive visualisation, the integration of the software in the care and work process and – where applicable – the possibility of interaction with the AIPA model.

It is **strongly recommended** that stakeholders – such as users and patients, clients or citizens – are involved in the design of the model presentation and explanation. **(4.1.1e)**

4.1.2 Information pertaining to the software

It is important for users of the software that they know and understand the characteristics of the software. The information requirements of the end-users are the determining factor for the way in which the characteristics of the software need to be explained clearly and unambiguously.

The aim of providing this information is to:

- Give the end-user insight into the intended purpose and the functioning of the software, thereby fostering trust;
- Give the end-user sufficient information to be able to explain the essence of the AIPA to third parties (e.g. the patient, client or citizen, if they are not the end-user), in other

words, to be able to interpret the meaning of the output in the intended medical context;

- Be able to substantiate the reliability of the software.

It is **strongly recommended** that digital instructions for use are drafted for the end-user, containing information about the use of the AIPA in the software. **(4.1.2a)**

It is recommended that the following aspects are included:

- Who the output of the AIPA in the software is intended for and in which medical context, e.g.:
    - Care organisation (non-care-related)
    - Care provider (between colleagues)
    - Care provider (in discussions with the patient, client)
    - Patient, client or citizen (in their own surroundings or elsewhere)
- In which manner (periodical) the information requirement of the end-user is and will be determined;
- The information requirement of the end-user and the provision of this information, e.g. by answering the following questions:
    - Care provider (and to a certain extent also patient, client and citizen):
        - Has this AIPA software already been implemented elsewhere?
        - What is the prediction of the AIPA based on? Which input variables have been used and which methodology has been used to process this data to form an AIPA? Where can I find information about training data that was used and the intended use (also refer to phase 1 through 3)?
        - Can dominant input variables be designated in relation to the predictions of the AIPA?
        - How certain are the predictions (also refer to phase 2 and 3)? Where can I find additional information about which validation data was used, which processes and methodology were followed and what the results were (also refer to phase 3)?
    - Patient, client or citizen:
        - What effect does the implementation of the AIPA software have on the process with my care provider?
        - What does this prediction entail for me as an individual?
        - Who can I contact if I have a question or complaint, or if I want to receive more information about the software or the AIPA?

- How certain can one be about the predictions (also refer to phase 2 and 3)?

4.2 Provisions for continuous monitoring

Continuous monitoring of the AIPA is an important aspect of quality management and a requirement stipulated by the MDR. For the design of a quality management system compliant with MDR, please refer to *ISO 13485 Medical devices - Quality management systems - Requirements for regulatory purposes.*

It is **strongly recommended** that a monitoring plan is drafted in this phase, so that the AIPA software can be tailored accordingly (refer to Section 6.2.1). **(4.2a)**

In the monitoring plan, the manufacturer distinguishes between their own information requirements and those of the care organisation.

It is **strongly recommended** that the option to perform monitoring of the data used, the model and the use of the AIPA after introduction in the practical situation (refer to phase 6) is facilitated in the software, so that at least the care organisation can make use of this option. **(4.2b)**

Deviation from this recommendation is permitted if this functionality does not offer any added benefit for the intended use.

It is **recommended** that an option is incorporated in the software to register whether the end-user actually follows the prediction (or classification, or treatment recommendation) made by the AIPA or not (and if not, why not), so that at least the care organisation can use this information. Also, pay attention here to – for example – the option to perform parallel implementation or testing. **(4.2c)**

Deviation from this recommendation is permitted if this functionality does not offer any added benefit for the intended use.

It is **strongly recommended** that input data is automatically validated in the software. **(4.2d)**

For example, by *out-of-domain detection*, where certain predictions that fall outside the domain are not presented to the end-user.

In addition, it is **strongly recommended** that monitoring be performed in the software for systematic shifts in the data. **(4.2e)**

For example, by detection of *data drift* in which the software automatically notifies the manufacturer and the software administrator when systematic shifts in the data appear to be taking place.

It is **mandatory** for the manufacturer of the AIPA software to facilitate feedback functionality as part of the quality management system, in the software if possible, for feedback from end-users and for reporting of technical problems. **(4.2f)**

The design and quantity of information requested in this feedback will vary per application and can differ for each phase of the development of the AIPA software. For example: in the initial phase, the option to perform a manual check and verification of predictions by the AIPA is strongly recommended.

4.3 Security

In general, the existing standards and regulations for software security are sufficient for AIPA software in the medical context, which is why we refer to the existing standards, regulations and guidelines for security of (medical) software, in particular:

- MDS2 statement: manufacturer disclosure statement for medical device security, ISO27002, NEN 7510.
- MDCG 2019-16 European Commission: Guidance on Cybersecurity for Medical Devices.
- UL 2900-1 ANSI/CAN/UL Standard for Software Cybersecurity for Network-Connectable Products.
- IMDRF/CYBER WG/N60 Principles and Practices for Medical Device Cybersecurity.
- FDA (most recent guidance from 2014 https://www.fda.gov/medical-devices/digital-health-center-excellence/cybersecurity, draft 2019).
- General Data Protection Regulation (GDPR).
- ISO/IEC TS 7110, ISO/IEC 27032, ISO/IEC 27014.

For AIPA software it is specifically important that input and output data will increasingly be stored and used in large quantities, particularly when re-training and re-calibrating models.

Therefore, it is **strongly recommended** that end-users who work with the software or the device that contains the AIPA are trained by the care organisation in terms of the use of secure (cloud) systems and the relevant standards and regulations in relation to the sharing, safety and privacy of data. **(4.3a)**

It is **mandatory** that version management is used for the software and the required and used training and test data sets for the development, validation and modification of the AIPA must be stored together with the corresponding version (also refer to phase 1 and 2). **(4.3b)**

4.4 Software testing

In general, the existing standards and regulations for software testing are sufficient for AIPA software in the medical context, which is why we refer to the existing standards, regulations and guidelines for testing of (medical) software. The guiding principle here is that complete traceability must be guaranteed from the translation of the *intended use* to the software requirements and design. This translation must subsequently be verified and validated. Please refer to the following existing standards and guidelines:

- IEC 62304 - Medical device software - Software life-cycle processes.
- IEC 82304-1 - Health software - Part 1: General requirements for product safety.
- IEC 62366-1 - Medical devices - Part 1: Application of usability engineering to medical devices.
- ISO 14971 - Medical devices - Application of risk management to medical devices.
- FDA, General principles of software validation, 2002.
- FDA, off-the-shelf software use in medical devices, 2019.

If components of the AIPA software have been developed by a third party, which is not managed by the manufacturer (also referred to as off-the-shelf or OTS software), then it is **strongly recommended** that these components are locally tested according to existing standards. **(4.4a)**

Refer, for example, to the FDA guideline from 2019 on off-the-shelf software use in medical devices.

In cases where the AIPA model itself can be designated as an off-the-shelf component, it is **strongly recommended** that phase 3 – Validation of the AIPA – be performed (once more). **(4.4b)**

4.5 References

1. Magrabi F, Ammenwerth E, McNair JB, et al. Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications. *Yearb Med Inform* 2019;28(1):128-34. doi: 10.1055/s-0039-1677903 [published Online First: 2019/04/26]
2. Cearns M, Hahn T, Baune BT. Recommendations and future directions for supervised machine learning in psychiatry. *Transl Psychiatry* 2019;9(1):271. doi: 10.1038/s41398-019-0607-2 [published Online First: 2019/10/24]
3. Kelly CJ, Karthikesalingam A, Suleyman M, et al. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019;17(1):195. doi: 10.1186/s12916-019-1426-2 [published Online First: 2019/10/31]

5 Impact assessment of the AIPA in combination with the software

Authors

Ilse Kant, Maarten van Smeden, Hine van Os, Teus Kappen, Jonas Teuwen, Leo Hovestadt, Ewout Steyerberg, René Drost, Sade Faneyte, René Verhaart, Lotty Hooft, Carl Moons, Niels Chavannes

Phase 5 covers the determination of the impact or added value of the use of the AIPA as part of the software on the envisaged medical practice or context, the medical treatment and the health outcomes respectively for the intended group (e.g. the patient, client or citizen). A Health technology assessment is also performed during this phase[2]. The manufacturer (or the developing care organisation, in the event of internal development) is responsible for determining the impact and added value. Still, this process is generally performed in collaboration with developers, care organisations and end-users. The impact or added value of the use of the AIPA – including the required software – on and in the medical practice can be achieved in different ways. For example by supporting the care provider, patient, client or citizen in making treatment or lifestyle decisions, or via an efficient (cost-effective) change in the care process.

At the moment, there are still relatively few AIPAs in use in daily medical practice and as a result, the impact has also been studied for relatively few AIPAs[3-6]. The integration of the AIPA in daily medical practice and care appears to be a stumbling block. It is hoped and expected that the implementation will increase in the coming years. For an overview of the current status regarding the implementation of AI in the medical context and practice, please refer to the report "Inventory of AI in health and care"[7].

5.1 Impact assessment and setting up accompanying study

It is **mandatory** for the manufacturer to perform an impact assessment of the AIPA (as part of the software) within the intended use. **(5.1a)**

This is necessary to test the potential added value of the AIPA in daily medical practice. In addition to the generic methods for empirical testing of the (added) value of (digital) innovations and prediction models in health care[8-13], several (AI-specific) steps will be discussed that are of particular importance to an AIPA[14-16].

It is **mandatory** that the development of the software that the AIPA forms part of and the accompanying impact assessment form a single process, in which the manufacturer ensures that end-users (e.g. care providers) and patients, clients or citizens are involved at the earliest possible stage and have several opportunities for contact. **(5.1b)**

If the impact assessment takes place in a care organisation, then an impact assessment can be considered as a form of implementation.

In the event that the impact assessment is (partially) performed within the care process – and thus implemented in the process – it is **strongly recommended** that an implementation plan be drafted as described in Section 6.1 and that relevant sections, such as the appointment of the implementation team and a local evaluation, a pilot or run-in period, be performed before a large scale empirical study is started. **(5.1c)**

It is **mandatory** that recommendation 5.1c is followed in the event of development within a care organisation. **(5.1d)**

Any required changes to the software can then still be implemented prior to the intervention period, as part of the impact assessment.

The impact assessment will be discussed according to a step-by-step plan. This step-by-step plan describes how to create an inventory of the expected effects of the use of software with an AIPA in medical practice, right through to the development of a comparative empirical study to demonstrate the expected effects compared to the current care (processes) in the envisaged context. This plan is based on the intended use of the AIPA (refer to phase 2) and the accompanying *indications* of *claims* that are made by the developer (refer to phase 4). The latter is mainly important to ensure compliance with the current legislation and regulations (MDR) and thus provide sufficient evidence of added value, which is required to be allowed to introduce the AIPA in medical practice.

The steps are as follows:

1. Expected effects: based on the intended use, list the effect that the AIPA is expected to have on the intended medical care process and health outcomes in the intended medical context;
2. Risk assessment: estimate potential risks and unintended effects prior to implementation of the AIPA in daily practice;
3. Human-machine interaction: Clarify how the care process and care provider will interact with the software before performing an empirical study.

5.1.1 The expected effects

It is **mandatory** to state clearly how the AIPA operates: independent or advisory, according to the *level of automation*; box 5.1. **(5.1.1a)**

The level of automation can affect the classification in terms of the previously mentioned Rule 11 in Appendix VIII of the MDR.

The intended use was recorded in phase 2, which will be included in digital instructions for use after development of the software, refer to Section 4.1.2.

In addition to the intended use that has been recorded, it is **mandatory** to record in more detail what the expected effects of the use of the AIPA are on possibly relevant (health and process) outcomes (in other words, define the *intended use* of the AIPA). **(5.1.1b)**

This determination can be divided into two layers: 1) The expected decisions by the end-user, based on the predicted outcomes or classifications; 2) the expected effects and

consequences of these predictions and decisions on later (health) outcomes of the patient, client, citizen and/or on the local care context or on society.

Box 5.1: level of automation

At the highest level of automation, an AIPA makes a prediction independently and makes a medical decision independently, in which the corresponding intervention is performed and is not checked by a human. This is virtually unheard of in clinical practice, although developments are ongoing in clinical practice, particularly in the field of radiological imaging, in which humans are no longer always involved in evaluating the radiological test images[1].

The *human-monitored* solutions are on the level below this. In this case, the software makes a prediction or classification independently, makes the next medical choice or decision and performs this independently, but has this checked – before or afterwards – by a human (usually a care provider or patient). If the human decision is made based on the recommendation of the software, we refer to this as *joint* or *computer-assisted decision making*. If *clinical decision support* is mentioned in a care context, then this usually refers to this level of automation.

Figure 1: the level of automation. The vast majority of AIPAs in software applications have until not been categorised as *computer assisted* or *joint decision-making*.

When determining the potential outcomes, it is **recommended** that the goals as defined in the *quadruple aim model* for the improvement of the care sector are taken into consideration, as applied in value-driven care[18]:

1) the improvement of the patient/citizen's experience regarding the quality of care;
2) the reduction of the care costs;
3) the improvement of the health of the population;
4) the improvement of the perception of the care provider. **(5.1.1c)**

Potential outcomes might include[12 14 17]:

- *Process outcomes and user-friendliness*
  Does the introduction of the AIPA result in immediate changes in the care process for the care provider (e.g. a faster or improved diagnosis or prognosis, potential for earlier hospital discharge or admission, improved work processes such as home monitoring or self-monitoring of the patient)?
- *Short-term health outcomes*
  Does the altered process after introduction of the AIPA result in improved health outcomes for the individual patients/citizens in the short term?
- *Long-term health outcomes*
  Does the altered process after introduction of the AIPA result in improved health outcomes for the individual patients/citizens in the long term, or in improved prevention (e.g. improved survival, higher quality of life, or improved daily functionality)?
- *Society*
  Does the altered process after introduction of the AIPA result in changes at the societal level (e.g. cost-effectiveness)?

Draw the expected chain of the medical care process that the software and the AIPA will form part of. Figure 2. serves as an illustration.

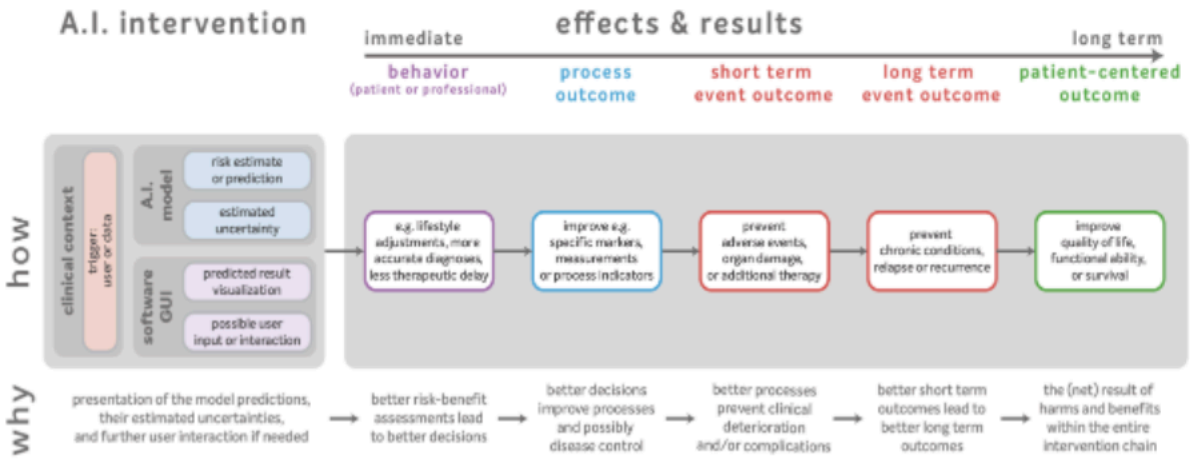

Figure 2: example of a drawing of the expected chain of events and the care or work process following the introduction of an AIPA in the envisaged medical context.

It is **strongly recommended** that the estimates of the expected effects as requested in 5.1.1b be made in a multidisciplinary setting, in consultation with the end-user(s) and patients, clients or citizens. **(5.1.1d)**

In this process, look at each step in the care chain of events that the AIPA will form part of, as set out in Figure 2. Take into consideration the *probability* (how great is the chance of this

effect occurring?) and the potential *consequences* (positive and negative). These estimates can be used to implement risk-mitigation measures (next step) and to set up a comparative empirical study.

5.1.2 Risk assessment

It is **mandatory** to perform a risk assessment to gain insight into the potential risks of the use of the AIPA in daily medical practice. **(5.1.2a)**

This includes the expected unintended decisions and effects in the entire care process (refer to Section 5.1.1) and reasonably foreseeable incorrect use;

It is **mandatory** to create an inventory of the potential undesirable effects (risks) of implementation of the AIPA in the care process per component of the care process in the risk assessment, in close cooperation with the stakeholders (e.g. end-users and patients). **(5.1.2b)**

It is **mandatory** to select and implement risk-mitigating measures for the risks identified during the risk assessment. **(5.1.2c)**

It is **mandatory** to include any sources of uncertainty as listed in 5.3.1a in the risk assessment. **(5.1.2d).**

It is **mandatory** to incorporate any risks identified in the risk assessment in the outcomes of the empirical study (refer to Section 5.1.4). **(5.1.2e)**

It is **strongly recommended** that stakeholders – such as users and patients, clients or citizens – are involved in the risk assessment. **(5.1.2f)**

The risk assessment by the manufacturer as described above is comparable to a prospective risk inventory (PRI) performed by a care organisation. The difference is that the risk assessment referred to above is a general assessment, covering the entire scope of potential use, as described in the intended use, instead of a PRI-targeted use of the AIPA in a specific care organisation. In the case of development by the care organisation, a PRI can be performed instead of the aforementioned assessment.

5.1.3 Human-machine interaction

The effectiveness of the interaction of the end-user with the software is of great importance to the impact of the AIPA software.

It is **mandatory** to ensure that the AIPA software interfaces with the current medical care processes and accompanying medical decision making as seamlessly as possible before an empirical study is performed (refer to Figure 2 in Section 5.1.1). **(5.1.3a)**

It is **mandatory** to involve several end-users in the local implementation team in order to achieve this (also refer to phase 6 for the further composition of an implementation team). **(5.1.3b)**

In addition, it is **mandatory** to create an inventory of the expected changes in the care context (e.g. changes in the work process) caused by the software, preferably in consultation with the intended user and patient, client or citizen. **(5.1.3c)**

Naturally, the relevant local medical guidelines must be followed. If new information has become available that demonstrates that following the current guidelines is no longer desirable as a result of the implementation of the AIPA (e.g. as a result of a new work process using the AIPA, which reduces the risk to patients or citizens), then this forms a reason to request a change in the local medical guidelines by the responsible organisation (e.g. a medical professional association).
It is **strongly recommended** that the desired presentation of the outcomes of the AIPA in the software (refer to phase 4) and the accompanying work process in which the AIPA is used (refer to Section 5.1.2) is demonstrated to the end-users before starting an empirical study, preferably in the form of a pilot, also refer to Recommendation 5.1c. **(5.1.3d)**

Consider a directive presentation, in which advice for action is given to the end-user[19].

It is **strongly recommended** that testing is performed to ensure that the design of the software, the required input and any actions required from the end-user are all consistent with the current workflow in order to encourage optimum use. For example, this should be consistent with the *work pressure* of the end-user. **(5.1.3e)**

It is **strongly recommended** to create an inventory of the so-called *facilitating factors and barriers* surrounding the implementation of the AIPA in practice, by using qualitative research methods such as focus groups and questionnaires. In addition to the developer of the AIPA and the manufacturer of the software, this should always involve several end-users and patients, clients or citizens[10 17]. **(5.1.3f)**

5.1.4 Comparative study

In order to ensure a valid quantification of the added benefit of the implementation of an AIPA in the context of the daily medical practices, it is **mandatory** to perform a comparative study, in which the (desirable and undesirable) effects of the use of the AIPA (refer to Section 5.1.1) are compared to a similar context, in which similar standard care is performed, without the use of the AIPA[9 11 12 17]. **(5.1.4a)**

The ideal design is a *randomised (adaptive) comparative study design* in which two groups are compared: one group will receive standard care (*the control group*) without the use of the

AIPA and one group in which the end-user will be informed of the recommendation made by the AIPA and can take action based on this recommendation in the care practice (*the intervention group*). In ideal circumstances, one would use a randomised design for this (randomised at individual or cluster level). Any deviation from this recommendation must be substantiated. Reasons for deviation include, but are not limited to: logistical infeasibility of randomisation, the risk of contamination of the control group, extended duration of a study (e.g. in the case of rare diseases or long-term effects), unfeasibly high costs, or an unfeasible number of patients to be included[9 11-13].

Alternatives to a randomised design are: a controlled prospective before/after design, interrupted time series, geographic comparison, or a cross-sectional randomised design with the treatment decision (instead of the individual health outcomes) as measures of effect.

It is **mandatory** to substantiate the choice of the population and context in which the AIPA software will be studied. **(5.1.4b)**

It is **strongly recommended** that a reasonably comparable population is selected for comparison with the target population for which the software was developed (phase 2). **(5.1.4c)**

The selection of the control group also forms part of the selection of this population. It is strongly recommended that a reasonably comparable population is selected for comparison with the intervention group, which is therefore also directly representative of the target population. If education is required for the intervention period (in other words: the use of the AIPA in practice), then a similar amount of education will need to be provided for the control group in order to prevent[10] the learning effect where possible.

Please refer to the SPIRIT-AI and CONSORT-AI statements for guidelines regarding the reporting of *randomized trials* in which artificial intelligence is used[20 21].

5.2 Health technology assessment

It is **strongly recommended** that phase 5 includes a model-based impact analysis, or a model-based Technology Assessment (HTA). **(5.2a)**

In other words, a mathematical model (e.g. a Markov model) is used to provide an objective analysis of the expected costs and benefits (added value) of the introduction of the AIPA in the medical practice compared to the current standard care as benchmark or control[2 22 23]. The result of such a HTA will become increasingly important in the approval of digital healthcare in the Netherlands and the EU. If reimbursement is essential, an appropriate HTA is thus required to be eligible for such reimbursement – or conditional reimbursement – of the implementation. The report “Valuable AI for health” includes a roadmap for performing an

HTA for AIPA software. This roadmap provides an overview of the costs and the potential funding sources of HTA research for AI and therefore also for an AIPA[24 25].

5.3 Uncertainty, risks and unexpected outcomes

5.3.1 Uncertainty in predictions

Source of uncertainty in implementation during phase 5 can include: the applicability of the AIPA in a different medical context than for which the AIPA model was originally developed (phase 2) or validated (phase 3), changes to the local care or work processes and a systematic change in the human-machine interaction as described in 5.1.1, Step 3.

It is **mandatory** that the manufacturer explicitly states which sources of uncertainty exist after performing the impact assessment and which mitigation measures have been implemented to minimise these uncertainties, which may be encountered during the introduction in the daily care practice. **(5.3.1a)**

The developer and end-users should pay particularly close attention to the transportability of the AIPA to a different medical setting and/or context[8 11 12 26].

5.3.2 Unexpected outcomes, vigilance

It is **mandatory** to log all unexpected outcomes that occur during the impact assessment and report these outcomes in accordance with the legislation and regulations. **(5.3.1b)**

Where applicable in the care context, the existing legislation and regulations regarding vigilance and the local safety management system[27] (SMS) must also be followed: MDR Article 5.5 (software developed in-house), MDR Article 10 (manufacturer's quality system), Covenant on Safe Implementation of Medical Technology in Medical Specialist Care (hereinafter referred to as: CMT)[28], MDR Article 80 (reporting of adverse events during clinical research), in the event of a product with CE marking: MDR Article 87 (safety reports). Manufacturers of an AIPA are obliged to implement a risk management system and a system for reporting incidents and corrective actions pertaining to safety in the field, both during the impact analysis (phase 5) and during and after implementation (phase 6). For the design of a quality management system compliant with MDR, please refer to *ISO 13485 Medical devices - Quality management systems - Requirements for regulatory purposes.*

5.4 References

1. Rodriguez-Ruiz A, Lang K, Gubern-Merida A, et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. *Eur Radiol* 2019;29(9):4825-32. doi: 10.1007/s00330-019-06186-9 [published Online First: 2019/04/18]
2. van Giessen A, Peters J, Wilcher B, et al. Systematic Review of Health Economic Impact Evaluations of Risk Prediction Models: Stop Developing, Start Evaluating. *Value Health* 2017;20(4):718-26. doi: 10.1016/j.jval.2017.01.001 [published Online First: 2017/04/15]
3. van Leeuwen KG, Schalekamp S, Rutten M, et al. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur Radiol* 2021;31(6):3797-804. doi: 10.1007/s00330-021-07892-z [published Online First: 2021/04/16]
4. Kelly CJ, Karthikesalingam A, Suleyman M, et al. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019;17(1):195. doi: 10.1186/s12916-019-1426-2 [published Online First: 2019/10/31]
5. Usher-Smith JA, Silarova B, Schuit E, et al. Impact of provision of cardiovascular disease risk estimates to healthcare professionals and patients: a systematic review. *BMJ Open* 2015;5(10):e008717. doi: 10.1136/bmjopen-2015-008717 [published Online First: 2015/10/28]
6. Kleinrouweler CE, Cheong-See FM, Collins GS, et al. Prognostic models in obstetrics: available, but far from applicable. *Am J Obstet Gynecol* 2016;214(1):79-90 e36. doi: 10.1016/j.ajog.2015.06.013 [published Online First: 2015/06/14]
7. KPMG. Rapport Inventarisatie AI in gezondheid en zorg in Nederland: KPMG, 2020.
8. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;21(1):128-38. doi: 10.1097/EDE.0b013e3181c30fb2 [published Online First: 2009/12/17]
9. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10(2):e1001381. doi: 10.1371/journal.pmed.1001381 [published Online First: 2013/02/09]
10. Kappen TH, van Klei WA, van Wolfswinkel L, et al. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. *Diagn Progn Res* 2018;2:11. doi: 10.1186/s41512-018-0033-6 [published Online First: 2019/05/17]

11. Moons KG, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012;98(9):691-8. doi: 10.1136/heartjnl-2011-301247 [published Online First: 2012/03/09]
12. Moons KG, Altman DG, Vergouwe Y, et al. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ* 2009;338:b606. doi: 10.1136/bmj.b606 [published Online First: 2009/06/09]
13. Riley RD, van der Windt DA, Croft P, et al. Prognosis research in healthcare: concepts, methods, and impact. Oxford, United Kingdom: Oxford University Press 2019.
14. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* 2020;368:l6927. doi: 10.1136/bmj.l6927 [published Online First: 2020/03/22]
15. Magrabi F, Ammenwerth E, McNair JB, et al. Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications. *Yearb Med Inform* 2019;28(1):128-34. doi: 10.1055/s-0039-1677903 [published Online First: 2019/04/26]
16. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019;25(9):1337-40. doi: 10.1038/s41591-019-0548-6 [published Online First: 2019/08/21]
17. Kappen TH, van Loon K, Kappen MA, et al. Barriers and facilitators perceived by physicians when using prediction models in practice. *J Clin Epidemiol* 2016;70:136-45. doi: 10.1016/j.jclinepi.2015.09.008 [published Online First: 2015/09/25]
18. Bodenheimer T, Sinsky C. From triple to quadruple aim: care of the patient requires care of the provider. *Ann Fam Med* 2014;12(6):573-6. doi: 10.1370/afm.1713 [published Online First: 2014/11/12]
19. Kappen TH, Vergouwe Y, van Wolfswinkel L, et al. Impact of adding therapeutic recommendations to risk assessments from a prediction model for postoperative nausea and vomiting. *Br J Anaesth* 2015;114(2):252-60. doi: 10.1093/bja/aeu321 [published Online First: 2014/10/03]
20. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med* 2020;26(9):1351-63. doi: 10.1038/s41591-020-1037-7 [published Online First: 2020/09/11]
21. Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* 2020;26(9):1364-74. doi: 10.1038/s41591-020-1034-x [published Online First: 2020/09/11]

22. Haverinen J, Keränen N, Falkenbach P, et al. Digi-HTA Health technology assessment framework for digital healthcare services. *Finnish Journal of eHealth and Welfare* 2019;11(4):326-41.
23. Jenniskens K, Lagerweij GR, Naaktgeboren CA, et al. Decision analytic modeling was useful to assess the impact of a prediction model on health outcomes before a randomized trial. *J Clin Epidemiol* 2019;115:106-15. doi: 10.1016/j.jclinepi.2019.07.010 [published Online First: 2019/07/23]
24. iMTA. Routekaart HTA onderzoek. Ministerie van Volksgezondheid, Welzijn en Sport: Erasmus University Rotterdam, 2021.
25. iMTA. Waardevolle AI voor gezondheid. Ministerie van Volksgezondheid, Welzijn en Sport: Erasmus University Rotterdam, 2021.
26. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;130(6):115-524.
27. VMS. VMS Handleiding Prospectieve Risico Inventarisatie.
28. (NVZ) NVvZ, (NFU) NFvUMC. Convenant Veilige Toepassing van Medische Technologie in de medisch specialistische zorg, 2e druk, 2016.

6 Implementation and use of the AIPA with software in daily practice

Authors

Ilse Kant, Maarten van Smeden, Hine van Os, Karen Wiegant, Laure Wynants, Anne de Hond, Bart Geerts, Nynke Breimer, Lysette Meuleman, Lotty Hooft, Carl Moons, Niels Chavannes

Phase 6 covers the implementation and use of the AIPA in healthcare provision. Central themes in this phase include implementation, monitoring and education by the care organisation. For an AIPA that forms part of a medical device as referred to in the MDR, phase 6 should also consider the legal requirements of *post-market surveillance* for manufacturers. It should be noted that the developer and the care organisation where the AIPA will be implemented can be one and the same party. In such cases, there will be no interaction between a manufacturer and a care organisation. In that case, the requirements and recommendations for the manufacturer should be interpreted as requirements and recommendations for the care organisation that develops the AIPA and any apparent duplications resulting from this situation can be ignored.

It is also possible that the AIPA will not be implemented as such by a care organisation, for example, when an AIPA will be used directly by patients, clients or citizens in their home situation, without the intervention of a care provider. In such cases, the sub-headings 6.1, 6.2.2 and 6.3.2 do not apply.

6.1 Implementation plan

When an AIPA is implemented and applied within a care organisation, it is **mandatory** that the care organisation drafts an *implementation plan*. **(6.1a).**

An implementation plan includes both the technical implementation of the AIPA and the software in the existing (IT) infrastructure and the embedding of the use of the AIPA in existing work processes. Please refer to the Covenant on Medical Technology[9] and the Guideline New Interventions in Clinical Practice[10] for a general guideline for implementation.

As part of the local implementation process, it is **mandatory** to evaluate the reliability and applicability of the AIPA by means of an assessment of (the results of) previous studies performed as part of phase 3 and 5. **(6.1b)**

If these results provide an inadequate indication of the reliability and applicability of the AIPA within the local context, additional validation of the AIPA can be performed (also refer to phase 3).

In addition, it is **mandatory** to introduce the AIPA – and the accompanying work process in which the AIPA is used – into the care process in a controlled manner, for example, in the form of a pilot, run-in period or by means of parallel implementation of the AIPA alongside the traditional care process. **(6.1c)**

It is **mandatory** to perform a prospective risk inventory (PRI) to gain insight into the potential risks of the use of the AIPA in daily medical practice. **(6.1d)**

A PRI explicitly states the medical relevance of every expected error and how likely it is that this error will occur. A PRI forms part of the *safety management system* of a specific care organisation. Please refer to the SMS manual Prospective Risk Inventory (SMS)[1] for execution of a PRI.

It is **strongly recommended** that the findings of the *impact assessment* and specifically the risk assessment as performed by the manufacturer in phase 5 – if available – are explicitly included in the PRI. **(6.1e)**

It is **mandatory** to evaluate the risks identified in the PRI and to ensure that these risks are then explicitly accepted or that risk-mitigating measures are selected. Both the acceptance of risks and any risk mitigating measures are included in the implementation plan. **(6.1f)**

It is **strongly recommended** that a data protection impact assessment (DPIA) is performed in the context of the GDPR, even if there is no legal or policy-related requirement for this. **(6.1g)**

It is **mandatory** for the implementation plan to include the envisaged implementation team. **(6.1h)**
It is important to bring together a specially appointed implementation team with a multidisciplinary background. This team should preferably consist of:

- At least two end-users, who should preferably be involved in the development of the AIPA from phase 1;
- A data scientist (or similar, e.g. clinical physicist, statistician, epidemiologist, biomedical technologist, technically specialised physician);
- IT specialist (with knowledge of the AIPA and software integration in existing systems);
- Project manager with a business background (or similar, e.g. a health scientist, policy-maker, administrator).

It is **strongly recommended** that the implementation plan be drafted in consultation with patients or clients. **(6.1i)**

It is **mandatory** that the implementation team is supported by the management of the care organisation in the ambition and considered choice for AI. They will take care of the policy regarding the following aspects: adequate resources for the IT department, protocols for application, reporting and monitoring, execution of the cost-effectiveness and impact assessment, any collaboration with a larger hospital and/or care organisation in the event of a lack of in-house knowledge, certification by the manufacturers, knowledge of buyers. **(6.1j)**

The managers involved can determine whether they are *AI-ready* in terms of all these aspects, for example by using available tools such as the *AI-readiness assessment* (https://www.ai-routekaart.nl/#readiness-assessment-intro).

6.2 Monitoring

6.2.1 Responsibilities of manufacturer or developing care organisation

It is **mandatory** for the manufacturer to monitor for technical errors in the AIPA and the accompanying software, incorrect use, incorrect predictions, fairness and unexpected adverse effects of the use of the software in daily practice. **(6.2.1a)**

This is important to guarantee safe and effective use of the AIPA in the long-term[2-7].

This is particularly important for the implementation of an AIPA, because some errors are unpredictable, hard to detect or new, or only occur over a period that was not studied in the impact assessment. Such errors can have a major impact on the care (and society) if the AIPA is used on a large scale[8].

There are two pathways for the developer and the manufacturer of the AIPA to ensure product safety: via a *post-market surveillance system* and via vigilance in terms of incident reports and safety warnings (Field Safety Notices). Please refer to the MDR and the CMT (MDR Article 87, Covenant on Medical Technology[9]) for further details about these systems. Again, please refer to the MDR for the design of a post-market surveillance (PMS) plan.

Additional to existing requirements of the PMS plan, it is **mandatory** at least to address the following:

- Monitoring and analysis of incorrect predictions by the AIPA (actual performance or outcomes differ from expected predictive performance and outcomes in terms of development and/or validation studies[8]). For example by monitoring of (depending on the type of application):
    - Miscalibration (e.g. the predicted prognosis is not accurate, the predicted probability of death due to cardiovascular conditions is over-estimated or under-estimated);
    - False-positive classification (e.g. the AIPA indicates that a tumour is malignant, but it turns out to be benign). To be measured using the positive predictive value (in %), preferably per relevant sub-group;
    - False-negative classification (e.g. the AIPA indicates that a tumour is benign, but it turns out to be malignant);
    - The error margin over time and the quality of data used.

- Monitoring for technical errors (e.g. the AIPA software does not produce output for certain individuals/patients, is not accessible, is not properly integrated in existing IT systems, the response time is not acceptable). The software must provide the option for the end-user to give feedback on poor performance of the software to the responsible administrator (e.g. the IT department) and/or the manufacturer (refer to phase 4).
- Monitoring of fairness: in order to measure fairness (and bias), it is mandatory to test whether no undesirable differences occur in the effects and outcomes of an AIPA for vulnerable groups as identified in phase 3, even after introduction of the AIPA in medical practice. It is important to collect as much data as possible about essential variables that quantify socio-economic determinants of health.
- Monitoring of the risks identified in the risk assessment (refer to Section 5.1.2) and analysis of uncertainties (refer to Section 5.3.1);
- Monitoring of *deployment bias, refer to Recommendation 6.2.1c*. **(6.2.1b)**

To establish *deployment bias*, it is **recommended**:

- To document automatically in the software whether the user follows the advice given by the algorithm or not (and why not). If this is not possible, provide a reason for this;
- To perform an evaluation regularly of the AIPA that is being used for the target group, which was also used for training and testing;
- To perform an evaluation regularly into the method employed by end-users to investigate whether the AIPA is used in accordance with the intended use. **(6.2.1c)**

6.2.2 Responsibilities of the care organisation

If an AIPA is being used within a care organisation, then this care organisation has a duty to monitor continuously for the correct functioning and use of the AIPA.

It is **mandatory** for a local monitoring plan to be drafted by the care organisation where the software that contains the AIPA will be implemented. **(6.2.2a)**

It is **mandatory** that the monitoring plan describes at least the following elements:

- Monitoring whether the intended and desired effect is achieved.
- Monitoring and analysis of incorrect predictions by the AIPA (actual performance or outcomes differ from expected predictive performance and outcomes in terms of development and/or validation studies[8]), for example in terms of miscalibration, false-positive and false-negative results (refer to monitoring and analysis of predictions of the AIPA in Requirement 6.2.1c for an explanation). This can be set up in collaboration with the manufacturer;

- Monitoring for technical errors (e.g. the AIPA software does not produce output for certain individuals/patients, is not accessible, is not properly integrated in existing IT systems, or the response time is not acceptable);
- Monitoring for unexpected effects for the care provider, patient, organisation, the care processes and/or society. These effects must be reported by the care provider, care organisation and/or the manufacturer;
- Model for incorrect use of the AIPA:
    - Automation (confirmation) bias: the predictions of an AIPA weigh too heavily in the decisions made by a care provider. Also monitor for *"deskilling"*: the loss of medical skills due to automation; and
    - *Deployment bias*: improper use (e.g. implementation for patients for whom the AIPA software has not been developed, failure to comply with model recommendations) or an incorrect interpretation of the outcome. Refer to Recommendation 6.2.2d.
- An analysis of the medical relevance for every expected error and how likely it is that this error will occur, based on the prospective risk inventory (PRI, refer to Requirement 6.1d). Monitoring for the risks identified in the PRI;
- Substantiation of which data will be collected for monitoring and how this is brought in line with end-users; and
- The frequency of monitoring and why this frequency was selected.
- Explicit reference of the obligation for the end-user to report to the manufacturer – and vice versa – any unexpected outcome for which the cause cannot be traced (compare to a *serious adverse event* (SAE) for an experimental treatment) **(6.2.2b)**

It is **recommended** as part of the monitoring plan to ask for the individual experiences of stake-holders (e.g. the patient and care provider) if possible. **(6.2.2c)**

To establish local deployment bias, it is **strongly recommended**:

- To perform permanent or periodic monitoring for connection of the software to care processes;
- To register how often the user uses the AIPA and whether a certain learning curve is present;
- To monitor the target groups: ensure that the AIPA for the preparation of medical interventions is being used for the target group, which was also used for training and testing;
- Perform an evaluation regularly into the method employed by end-users to ensure correct use and guarantee meaningful interpretation of the results;

- Check data entry and storage for possible measurement errors, even after implementation and commissioning of the algorithm. **(6.2.2d)**

6.3 Education

6.3.1 End-user

It is **mandatory** that the end-user (e.g. a patient or the care provider) has access to information about the topics described in Box 6.1, to be supplied by the developer or manufacturer. **(6.3.1a)**

If the end-user is a care provider, it is **mandatory** that the care provider has access to education about the topics described in Box 6.2. **(6.3.1b)**

This aims to empower the care provider to use the AIPA in a competent manner in medical practice.

It is **strongly recommended** that this education be repeated at regular intervals, depending on the application and the medical context. **(6.3.1c)**

This can be supported by supplying the end-user with information (by the manufacturer or developer) about the following: the results of the validation and impact assessment, as performed in phases 3 and 5, the population on which the training and validation was based (phases 2 and 3), the performance of the AIPA (phase 3), the expected error margin, statistical knowledge and explicability of the algorithm.

Box 6.1: Education of the end-user

The end-user must have access to education regarding the following points:

- The intended use of the specific AIPA, limitations of this (intended) use and the accompanying user manual, as mandated by the MDR.
- Interpretation of the outcomes of the AIPA.
- The potential errors of use that can be made, as described in Section 6.1. Special attention should be paid to the transportability and generalisability of the AIPA to the local medical environment. Ensure that the AIPA can set its own boundaries, by issuing a warning when a data point deviates too far from the training population, by not providing a prediction in certain cases, or by providing confidence intervals (also refer to phase 4) and training end-users in the interpretation of this.
- The potential benefits that can be achieved by implementation of the AIPA.
- Instructions about (definitions and quality of) data entry that is expected of the end-user.

6.3.2 Care organisation

It is **mandatory** that the care organisation has access to information and/or education about the topics described in Box 6.2, to be supplied by the developer or manufacturer. **(6.3.2a)**

The purpose of this is to make the care organisation *aware* of the responsibilities associated with the purchase or in-house development of an AIPA that will be implemented and used in medical practice. The responsibility and duty to obtain information rests with the care organisation, it is a pre-condition to have or gain the required knowledge for the implementation of AIPAs.

It is **mandatory** that end-users (i.e. care providers) are granted the time and opportunity for education regarding the AIPA. **(6.3.2b)**

Box 6.2: Education of the care organisation

- Legislation applicable to the AIPA (including MDR, also refer to *Phase 4*).
- Required technical administration. This includes (but is not limited to): cloud management, support of IT infrastructure, storage and service of large data flows, communication with the administrator of electronic patient dossiers.
- The results of the evaluation and assessment performed as part of phase 3 and 5.
- The effect of AI on the care institution: attention must be paid to the change in work processes and autonomy of the care providers.
- Costs/benefits: development of a business case, which includes not only the costs of purchasing, but also the costs of required activities in terms of education, implementation and management.
- Ethical considerations (refer to Fairness and algorithmic bias, phase 3)

6.4 Rights and Duties

The stakeholders involved in the implementation and use of AIPA software in medical practice each have *rights* and *obligations* resulting from this guideline to ensure a responsible introduction, in other words, taking into consideration the safety and added value of AIPA software in practice. A distinction has been made between four stakeholders: the care provider, the care organisation, the patient and the manufacturer.

Use of the list of rights and obligations of care providers, care organisation, patient/citizen and manufacturer to check and determine that all rights and obligations can be exercised effectively is **strongly recommended**. **(6.4a)**.

6.4.1 Care provider

Rights:

- To receive support in knowledge about specific AIPA by care organisation and manufacturer, comprehensible end-user information and training for use and monitoring;
- Transparent communication (by manufacturer of the AIPA software and/or third parties who performed studies for validation purposes) about previous studies;

- Feedback on reported incidents.

Obligations:

- To be consciously competent in the use of the AIPA;
- To comply with the intended use in the interests of the patient and manufacturer;
- To perform implementation according to the implementation plan of the care organisation;
- To ensure transparency about the use of AI in prognosis or diagnosis (to care organisation and patient, for example in the file);
- To provide feedback on incidents in the quality management system;
- To ensure transparency to the patient about the use of patient data for improvement of AI and to obtain informed consent where necessary.

6.4.2 Care organisation

Rights, in the situation where an AIPA produced by a manufacturer is implemented:

- To be supported by the manufacturer in obtaining knowledge about the specific AIPA;
- Transparent communication (by the manufacturer) about previous studies;
- Feedback on reported incidents;
- Quality management system that facilitates reports (provided by the manufacturer).

Obligations, both when implementing an AIPA produced by a manufacturer and in-house development:

- To provide support to employees who work with the AIPA;
- To ensure proficiency of (AI) involved employees;
- To comply with the intended use in the interests of the patient and the manufacturer;
- In general, state that AI is being used;
- Feedback on incidents (to the manufacturer, care organisation, patient and in file);
- Ensure that an implementation plan is in place, that will be executed in line with the guideline (refer to Section 6.1);
- To be transparent about the use of patient data for improvement of the AIPA;
- Work with a quality management system.

6.4.3 Patient, client or citizen

Rights

- The intended use of the AIPA is observed by the care provider and care organisation;
- Transparency about the use of AI in the care organisation;
- Transparency about the use of patient data for improvement of AI;

- Support from the manufacturer in the event of a direct relationship for self-care applications (e.g. *e-health* applications);
- Feedback by manufacturer on individual incidents reported by the patient;
- Feedback on bugs and quality control;
- The right to stop sharing data at any time. The right to be forgotten (as referred to in GDPR Article 17);
- *Possibly: t*ransparency about information about the AIPA used (e.g. results of previous studies).

Obligations

- If consent is provided, ensure correct entry/supply of data, as long as the AIPA software is used;
- Feedback on incidents (to care provider / care organisation or directly to the manufacturer);
- Comply with intended use (in case of direct relationship, e.g. correct data entry and data measurement at agreed times).

6.4.4 Manufacturer or developing care organisation

Rights

- Obtain information about reports from the care organisation, care provider or patient.

Obligations

- Support to care provider and patient for maintenance and knowledge of correct use of the AIPA (as mandated by the MDR);
- Transparent communication (to care organisation/end-user) about previously performed studies;
- Supply of information about monitoring and quality management system;

Processing and feedback to care provider or patient on incidents reported by either party (including obligations regarding vigilance, as discussed in 5.3.2).

6.5 References

1. VMS. VMS Handleiding Prospectieve Risico Inventarisatie.

2. Buruk B, Ekmekci PE, Arda B. A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Med Health Care Philos* 2020;23(3):387-99. doi: 10.1007/s11019-020-09948-1 [published Online First: 2020/04/03]

3. Floridi L, Cowls J, King TC, et al. How to Design AI for Social Good: Seven Essential Factors. *Sci Eng Ethics* 2020;26(3):1771-96. doi: 10.1007/s11948-020-00213-5 [published Online First: 2020/04/05]

4. Kelly CJ, Karthikesalingam A, Suleyman M, et al. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019;17(1):195. doi: 10.1186/s12916-019-1426-2 [published Online First: 2019/10/31]

5. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019;25(9):1337-40. doi: 10.1038/s41591-019-0548-6 [published Online First: 2019/08/21]

6. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* 2020;368:l6927. doi: 10.1136/bmj.l6927 [published Online First: 2020/03/22]

7. Magrabi F, Ammenwerth E, McNair JB, et al. Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications. *Yearb Med Inform* 2019;28(1):128-34. doi: 10.1055/s-0039-1677903 [published Online First: 2019/04/26]

8. Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* 2020;26(9):1364-74. doi: 10.1038/s41591-020-1034-x [published Online First: 2020/09/11]

9. (NVZ) NVvZ, (NFU) NFvUMC. Convenant Veilige Toepassing van Medische Technologie in de medisch specialistische zorg, 2e druk, 2016.

10. Zorginstituut Nederland, Federatie Medische Specialisten (voorheen OMS). Leidraad Nieuwe Interventies in de Klinische Praktijk. https://www.demedischspecialist.nl/sites/default/files/Leidraad%20Nieuwe%20interventies%20in%20de%20klinische%20praktijk%20def.pdf. Published Oktober 2014. Accessed December 8, 2021

Future perspectives

Authors

Authors: Ilse Kant, Maarten van Smeden, on behalf of the medical AI working parties

The objective of the working parties “guideline medical AI” was to draft a joint, publicly accessible set of criteria for evaluation and testing of medical AIPA software, as commissioned by the *Ministry of Health, Welfare and Sport*. Considering the rapid developments in the field, this guideline should be viewed as a dynamic standard. For example, the working party members have already listed several generic topics and future developments and they deem it important that these aspects receive special attention in future versions.

Dynamic updates of AIPAs

AI technology is developing so rapidly that the current standards and regulations may be outpaced. The working party members conclude that compliance with the MDR in its current format is either very uncertain or very cumbersome for an AIPA that needs to be re-trained very frequently using new data in order to ensure continuous improvement and updating of the AIPA whilst in use. In addition, the uncertainty amongst developers about the application of the MDR in its current form inhibits many potential (future) AI applications. This topic was not discussed in the current version of the guideline, so as not to create false expectations about compliance. It would be highly desirable to include this topic – and specifically the evaluation of these types of updates – in future versions of the standard.

Cost-effectiveness evaluations

The current version of the standard does not take into consideration market authorisation and the route to inclusion of AIPA in healthcare packages and reimbursement by healthcare insurers. The outcomes of impact analyses using model-based HTA analysis of the AIPA (refer to phase 5) will increasingly play a role in the market authorisation and the scope for inclusion of AI in healthcare packages of insurers. Therefore, this aspect will need to be included in the guideline in future.

Sharing data

The working party members wish to emphasise the importance of (inter)national collaboration in the field of data sharing and AIPAs, in order to encourage and support future developments in healthcare. Examples of initiatives that are already working on this are the NICE foundation[1] and the NEED foundation[2] databases for ICU and A&E data, respectively, Health-RI and the Personal Health Train[3]. The Dutch AI coalition working party *data sharing* has published a guideline for handling data in collaborations[4]. An example of a new development that can be implemented in this type of collaboration is *federated learning*. Federated learning facilitates large-scale collaboration between multiple institutes, without the use of a central database and is therefore very privacy-friendly. This topic has not been

included in the current version of the guideline. It would be desirable to include this topic in future versions of the standard.

Early multi-disciplinary work

The working party members perceived various stumbling blocks that could hamper the implementation of an AIPA in medical practice. The number of applications is increasing rapidly in the specialised medical field, particularly in radiology. However, many projects stall in the early phases (phase 2 or 3 of the guideline). The working party members regularly mention the early involvement of several end-users of the AIPA as a solution. Other stumbling blocks perceived by the working party members relate to the transportability and generalisability of predictive models to a different medical context (e.g. a different hospital, country or even medical work process). Much is still unknown about this. More research is needed to make more realistic estimates (beforehand) about if – and when – an AIPA should be re-validated and/or re-trained in a different context before being implemented. Therefore, no specific recommendations about this were included in the current version of the guideline. In addition, it would be desirable to include specific recommendations about managing an expansion of the intended use of an existing AIPA.

Monitoring in practice

Phase 6.1 of the guideline describes monitoring after importation of AIPA Software in the *real world* (in other words: beyond the study setting) in medical practice. The working party members conclude that – following implementation of AIPA software in practice, where a treatment decision is usually linked to the outcome of the AIPA (e.g. for decision support) – the monitoring of *miscalibration* based on data from the practice appears to be problematic. Much is still unknown about this issue and further research is desirable. It would be desirable to provide greater clarity on this topic in future versions of the guideline.

Education

The rapid emergence of digital healthcare could result in disruptive changes to medical care processes. However, the working party members noticed a general lack of basic knowledge about AI in care organisations and on the shop floor, in other words, amongst doctors, nurses and others who will increasingly be confronted with AI in their daily care activities. In some cases, even the patient should have this basic knowledge. In order to introduce AIPA technology effectively and responsibly in daily medical practice, this knowledge deficit will need to be resolved, firstly in the education of care professionals who are currently employed and secondly in the training of new care professionals. In addition, developers and manufacturers will need to establish basic knowledge in the field of safe and effective introduction of an AIPA, in which this guideline can serve as a guideline. The further

development of this standard should continue to focus on (future) manageability and feasibility of the standard for developers and manufacturers.

References

1. Stichting NICE [Available from: https://www.stichting-nice.nl/dd/#start.
2. Stichting NEED [Available from: https://www.stichting-need.nl/.
3. Health-RI [Available from: https://www.health-ri.nl.
4. Klauw Kvd, Bastiaansen H, Ette Fv. Verantwoord datadelen voor AI: Nederlandse AI coalitie, 2020.

Colofon

This guidance was created by a broad representation of experts in healthcare, supported by the action team within the programme *waardevolle AI voor gezondheid* (valuable AI in healthcare) of the Dutch ministry of Health, Welfare and Sport. This guideline provides a description of what the work field considers good professional conduct in the development, testing and implementation of an Artificial Intelligence Prediction Algorithm (AIPA) in the medical sector, including public healthcare. The ambition is to make this guideline well-accepted so that it can be used as a normative document in the future.

If you have any inquires or remarks, please reach out to one of the main authors:

dr. Maarten van Smeden - M.vanSmeden@umcutrecht.nl

dr. Ilse Kant – I.M.J.Kant@lumc.nl

Waardevolle AI