Physica Medica 83 (2021) 257–263

Contents lists available at ScienceDirect

Physica Medica

journal homepage: www.elsevier.com/locate/ejmp

Original paper

Procurement, commissioning and QA of AI based solutions: An MPE's perspective on introducing AI in clinical practice

Hilde Bosmans [a,1,*], Federica Zanca [b,1], Frederik Gelaude [a]

[a] University Hospitals of the KU Leuven, Leuven, Belgium
[b] Palindromo Consulting, Leuven, Belgium

ARTICLE INFO

Keywords:
Artificial intelligence
Procurement
Commissioning
Virtual clinical trials

ABSTRACT

Purpose: In this study, we propose a framework to help the MPE take up a unique and important role at the introduction of AI solutions in clinical practice, and more in particular at procurement, acceptance, commissioning and QA.

Material and methods: The steps for the introduction of Medical Radiological Equipment in a hospital setting were extrapolated to AI tools. Literature review and in-house experience was added to prepare similar, yet dedicated test methods.

Results: Procurement starts from the clinical cases to be solved and is usually a complex process with many stakeholders and possibly many candidate AI solutions. Specific KPIs and metrics need to be defined. Acceptance testing follows, to verify the installation and test for critical exams. Commissioning should test the suitability of the AI tool for the intended use in the local institution. Results may be predicted from peer reviewed papers that treat representative populations. If not available, local data sets can be prepared to assess the KPIs, or 'virtual clinical trials' could be used to create large, simulated test data sets. Quality assurance must be performed periodically to verify if KPIs are stable, especially if the software is upscaled or upgraded, and as soon as self-learning AI tools would enter the medical practice.

Discussion: MPEs are well placed to bridge between manufacturer and medical team and help from procurement up to reporting to the management board. More work is needed to establish consolidated test protocols.

1. Introduction

There is currently intense interest in the application of Artificial Intelligence (AI) based methods to the field of medical imaging. Progress is rapid, yielding many exciting developments, yet a number of steps are required before a new AI solution successfully enters routine medical practice, and more specifically in radiology, nuclear medicine and radiotherapy. Medical physics experts (MPEs), with the correct training, are ideally placed to help ensure the timely and appropriate application of any new AI supported application. This the topic of the current paper.

The number of commercially available AI applications is increasing [1]. This can be clearly seen from the number of scientific publications and presentations at medical conferences, the number of commercial booths at these conferences and the many webinars being organized on the same topics. In parallel to this, larger institutions may also develop numerous in-house AI solutions. Promises in terms of improved patient outcome, harmonized workflow, availability of quantitative data and other fruitful applications of AI are appealing [2]. However, dealing with AI solutions from the procurement stage, through commissioning and final integration within clinical practice poses several challenges. This is especially the case if the product is very new or even unique, if the stated technological advance is extensively documented or described in peer reviewed publications and if hospital managers require, prior to any purchase order, a complete documentation of the expected benefits and costs. The medical team can be overwhelmed by the many clinical, ethical, economical and technical questions to be answered prior to the clinical implementation of an AI package. In addition, the specific aspects of AI solutions may go beyond the expertise of technical purchase departments in the hospital and even beyond the primary field of interest of the PACS managers.

MPEs can form the necessary link between the different stakeholders involved in the decision to clinically integrate an AI based software

* Corresponding author.
E-mail addresses: hilde.bosmans@uzleuven.be (H. Bosmans), federica.zanca@palindromo.consulting (F. Zanca), frederik.gelaude@uzleuven.be (F. Gelaude).
[1] Contributed equally to this work.

https://doi.org/10.1016/j.ejmp.2021.04.006
Received 12 January 2021; Received in revised form 24 March 2021; Accepted 6 April 2021
Available online 10 May 2021

solution as they are familiar with assessing both technological aspects as well as the clinical applicability and the quality of devices in diagnostic or radiotherapy/oncology departments. MPEs are also used to participating and providing support during procurement and commissioning. They write and evaluate technical specifications and perform acceptance testing and commissioning prior to the clinical use of some medical device. Most importantly, they possess technical-medical communication skills [3]. Some of MPEs also have expertise in data collection or database creation and in testing interoperability across several IT tools. In order to embrace AI tools, in the ideal setting, a team of experts that embraces AI methods is needed, with activities evenly balanced between team members according to their particular competences. Integration of AI may be a time consuming process.

It is, however, not yet clear among the clinical and scientific community how the procurement or commissioning of AI tools should proceed. The role the MPE is clearly stated in the European Medical Exposure Directive of Dec 2013 [4]: "Member States shall ensure that the optimization includes the selection of equipment, the consistent production of adequate diagnostic information or therapeutic outcomes, the practical aspects of medical radiological procedures, quality assurance, taking into account economic and societal factors." Many Member States have in the meantime appointed this task to the MPE, to be performed in close cooperation with the medical teams. Many AI based solutions add to the diagnostic information or therapeutic outcome and are therefore included. For example, in Belgium, the implementation of the Medical Exposure Directive [5] places the hardware and software components that constitute medical-radiological devices on the same level and therefore at the core of MPE activities. In a white paper on behalf of the European Federation of Organisations of Medical Physics (EFOMP) [6], M. Kortesniemi is clear: "Quantitative aspects of data validation, quality control, physically meaningful measures, parameter connections and system modelling for the future AI methods are positioned firmly in the field of the medical physics profession". This is also stated in the EFOMP statement Policy [3].

Document RP162 from the European Commission [7] gives an active role to the physicist with respect to new technology: "the MPE should agree suspension levels with the holder. The levels proposed by the MPE must be framed to be effective for new technology, take account of related longer established technologies, any CENELEC/IEC standards available, newly available test methods, the manufacturer's recommendations, related scientific and professional opinion/published literature and the maxim that the new technology should aspire to be at least as safe as the technology it is replacing." It is true that quality assurance and risk management will be different from what most MPEs are familiar with: assessing the quality of AI based software, from procurement through to routine use, possibly with upscaling and upgrading of software through its lifecycle lacks specific guidance. Many skills have yet to be developed and there are no clear-cut test protocols, a common tool in the MPE's daily practice.

The purpose of this text is to position the MPE within the hospital workspace at a time when AI methods are being assimilated in radiology, nuclear medicine or radiotherapy departments by providing a framework for the different aspects involved in this process. The focus of the paper will be on procurement, acceptance testing, commissioning and quality assurance.

2. Material and methods

To identify the role that MPEs can play with respect to the introduction of AI applications in clinical practice, we have extended the steps necessary for the introduction of Medical Radiological Equipment in a hospital setting [7] to cover AI applications (see Table 1). The next sections follow the chronological steps described in the table. Based on preliminary, personal experience, we provide examples and suggestions on how to implement each step in clinical practice.

Table 1
Summary of the chronological steps in the adoption of an AI tool in healthcare, .

Terms	Scope
Procurement	To guide the selection of the optimal AI application in terms of safety, performance, match with the target use case, usability, ethical aspects and price.
Acceptance testing	To ensure compliance of a new AI application with its safety and performance specification at installation.
Commissioning	To prepare the AI application for clinical use and the roll-out within the local clinical workflow.
Quality assurance	To assure that the AI application operates over time as expected, for its purpose.

adopted from [7]

3. Results

3.1. Procurement

The first step towards the procurement of an AI system is the identification of the clinical need and the gap between the desired end result and the current process, as well as the documentation of a potential solution using an AI application. The aims could be purely technical or technological [8–12] with an indirect impact on health care, but many of the commercially available tools aim directly at improving patient outcome, clinical efficiency and/or operational efficiency [13–15], for example.

Procurement is a vital step and therefore a multi-disciplinary team composed of different stakeholders with the right skills should be put in place from the start. This will involve, amongst others, representatives of the medical team, the MPE, the purchasing manager, the IT manager, the data scientist, the ethics and data protection experts, and the final user.

Software employing Artificial Intelligence is generally classified as a Medical Device under the new Medical Device Regulation (MDR) [16] and therefore belongs to a specific class as a function of the patient condition (non-serious to critical) and the significance of information provided by the software to the healthcare decision (inform clinical management, drive clinical management, treat or diagnose). As such they are designed, implemented and tested using quality standards. Note also that manufacturers are asked, under the MDR, to identify potential risks associated with the use of the software and associated mitigation measures. All users have to adhere to the General Data Protection Rules (GDPR) [17].

The team in charge of procurement should at first define which metrics will be used to compare different solutions available on the market or developed in house, and which local Key Performance Indicators (KPIs) will be used to select the device that accomplishes the local needs best. The definition of KPIs does not just relate to safe use but also to the verification that the AI application meets specified performance levels when data and images from local practice are used. Hence, data from the literature or from the package documentation may not fulfill the KPI.

These indicators will define the technical specifications to be added in the tender document. While tendering is common practice for classical medical-radiological equipment, typical standard operating procedures (SOPs) may not exist for new AI tools. Both metrics and KPIs should start from the 'intended use' and the 'indications for use' of the software product. The MPE (together with the clinical team) has to address the aspects specific to local use, including the identification and assessment of possible risks associated with the device, including incorrect use. In this respect the MPE can start from the documentation provided by the developers, identify potential local 'risk scenarios' and test the means taken to minimize the risks, in compliance with ISO standards. Local KPIs however go beyond the standard safety aspects. KPIs may vary substantially from application to application and include processing and reconstruction, image quality enhancement, organ

segmentation, lesion detection, lesion classification, monitoring, prediction, support for therapy and even far beyond all this in multidisciplinary applications. The AI manufacturer might report performance in terms of area under the ROC [18] curve for a classification tool or the amount of time saved if the aim is a better workflow or process. The assessment of such performance is however not straightforward, especially for new software for which there is no standard reference, like in new therapy applications.

Once the KPIs have been identified, the quality of the AI software – in terms of data and knowledge that was used to build the software and the clinical evidence that was generated by the manufacturer – should be assessed. The databases that have been used should be representative, unbiased and sufficient to yield robust results in actual practice. Linked to this aspect, it is important to notice that the local situation may be different from the data sample used during the manufacturer's design and development process and the impact of these differences should be investigated. As companies may consider this as confidential, it would be helpful if regulatory instances would make the release of some of these aspects a requirement or create independent reference datasets for performance testing of an AI application.

MPEs involved in the assessment should study the documentation of the product and understand any limitations of the product that may arise from how the software was designed, trained and tested. They should carefully help balance the technical challenges and the benefits for patient and hospital, and work towards a team decision. It is crucial to select a software tool that fits local needs and is compatible with the local input data. Manufacturers need to be clear about the expertise required for using the system and the purchaser might want to assess if the final users need extra training for software utilization.

Then, a cost/benefit analysis should be considered [19,20]. Companies developing AI software for radiology applications usually work within pay-per-analysis models or with a subscription fee for an entire site, per workstation, per user, or per study license. The latter can be beneficial for low volume examinations. Prices vary widely depending on customer need and most often will not include costs for hardware, installation, training or maintenance. While the aim of these software tools is to bring about clinical or operational improvements, this alone may not be sufficient to drive software implementation unless direct cost savings or operational time reduction can also be attributed to the software. It is even more difficult to assess upfront what will be the impact on job satisfaction or their use in potential scientific studies. Furthermore, benefits may go beyond a single department, embracing surgery, digital pathology and genomics, or even beyond the walls of the hospital such as in breast cancer screening networks. It is also possible that one hospital explores a software tool on behalf of a cluster of hospitals or a regional association of purchasing departments / hospitals or even the local government [21]. The initial contracts between the manufacturer and the purchasing team may have to foresee such upscaling. The assessment should then cover multi-stakeholders – and eventually – incorporate hard decision criteria.

Finally, there are particular aspects associated with AI that require attention from day one in the contract negotiation. Some examples include: how to proceed if the AI solution would lead to interesting medical findings beyond the intended use? Some AI tools learn continuously, especially if used on many cases. Can the manufacturer support a learning algorithm on the long run? Is the hospital willing and allowed to share specific input images with the manufacturer? It must be stated that the use of a medical device beyond its intended use will excuse the company from liability.

Given these points, the decision making at the procurement stage can easily become complex and time consuming, yet is essential if the correct choices are to be made.

3.2. Acceptance testing

In the structure familiar to MPEs, and in line with legislation, procurement would be followed by acceptance tests that are performed to verify whether the correct system has been successfully installed (Table 1), in line with specifications made at the procurement stage and whether national/regional acceptability criteria have been met [4]. The radiology manager must ensure that new equipment is not used clinically until a critical examination and a commissioning test have been completed. Acceptance tests may have to be performed a few times: after the implementation of a trial version, then when the complete AI tool is installed and when a tool gets upgraded or upscaled over different centers.

Acceptance testing should cover different aspects:

- The package specified in the procurement document has been supplied, installed and tested following the manufacturer's instructions to establish that the package is functioning as designed
- The installation of the device and the correct functioning of the application programming interface (API). The latter comprises the type of requests that can be made, how to make them, the data formats that should be used, etc.
- Training by application specialists on the utilization of the software as part of the existing clinical workflow.
- The trainees should be representative of the intended end-user population.
- The checking of consistency and repeatability of the output.
- The selection of test cases reflecting critical examinations, including borderline cases for the intended use of the system. An examination can be considered critical from a number of perspectives, from either a test whether the software in combination with the local computer infrastructure can handle the throughput; or a compatibility test with data formats and types of input data; or to verify whether data are correctly integrated to the appropriate patient data files; or whether the data is available for other analyses such as big data applications and data mining, etc.
- The typical error scenarios.
- Unexpected or incomplete data should also be tested for appropriate output. Possibly, cases as described in the risk management analysis for the CE marking could be considered and the mitigation actions tested. Ideally the system is also expected to behave consistently under these critical circumstances.

On successful completion of the acceptance tests, the acceptance report will be signed and the payment for the device approved.

3.3. Commissioning

Commissioning ensures that the equipment is fit for clinical use and baseline values are acquired for future routine performance testing. Contrary to x-ray devices and as far as we know, there are no standard test protocols available for these tasks. Without doubt, safety assessment should target dangers other than 'too high patient dose' and focus on issues such as a wrong or missed diagnosis, incorrect classification, analyses of reduced quality and incorrect quantitative results.

Commissioning tests should go to the core of the performance and test the suitability of the AI tool for the intended use in the local situation. A number of publications [13,13b] suggest how MPEs should approach such tasks: "it requires the medical physics 3.0 philosophy, i.e. find significant quantifiable indications and test how the device is creating a positive clinical change. It is important that the MPEs apply a healthy – though not paralyzing – dose of skepticism and rigor to validate the model. Intrinsic to the MPE expertise is not to fall in the trap of "believing everything they are told."" The investigation should start from the intended use declared by the vendor, or from the requirements put forward in the original plan or during the contract negotiation phase. Some AI tools have sensitivity and specificity as an input parameter that can be adapted to the local needs. The critical cases at acceptance testing are used to ensure compliance of the new AI tool with

its safety and performance specification as provided by the manufacturer in the supplied documentation. At the commissioning stage, the AI application is setup for use within the clinical workflow (and in fact may change the workflow). It may be possible to test this with the same database. A similar process should be applied following updates or upgrades, possibly using the same datasets.

We propose 3 potential approaches to commissioning.

Search the literature for peer reviewed publications that have been performed with exactly the same AI software application.

This could include verification of the scientific literature if a similar dataset is available, or a visit to a clinical site that has implemented the AI package. Extrapolation of these results to the local situation is possible if input data are similar and if the specifications put for the results are similar. MPEs can help judge the first aspect in particular: the technical specifications of the input images must be similar before the results of any AI tool can be extrapolated.

Advantages: if large clinical studies are available or if regulatory institutions such as the FDA have given clearance for the particular use of the software, the decision to start working with the software may be smoother, faster and easily accepted by ethical committee and referral doctors.

Limitations: Literature on testing software applications is extremely scarce to date. This may apply especially for local software solutions which follow specific legal rules within the EU [16]. The number of studies with results relevant to the local expected performance may be very limited. Moreover, one should make sure that the reported literature, which could be relevant for the local commissioning of the same AI tool, follows well recognized standards for study reporting, such as TRIPOD-ML, SPIRIT-AI and CONSORT-AI [24–26].

Example: Publication [27] is a comparative study on deep learning algorithms for the detection of lymph node metastases in women with breast cancer. The comparison was organized as a challenge and results are given for many algorithms in terms of area under the ROC curves. The patient group considered in the paper may be similar to the local situation.

Create a database with local cases for which the gold standard result is available or run a pilot where the software is used in parallel with current workflow.

In the absence of suitable publications or relevant documentation, a next step could be to run a local clinical trial. When a number of candidate AI tools are being compared then such studies must be carefully set up. A few requirements apply: (1) the dataset should be sufficiently large to obtain meaningful results, (2) the dataset should also include rare cases put forward by the medical team, and (3) the datasets should not have been used to create or train the AI tools, and this might mean the use of local rather than publicly available databases. MPEs should be able to judge whether the quality levels in the test data correspond to local practice.

It may be good practice to construct a local database and keep this database separate for local testing purposes only. The construction of such a database may be straightforward if structured reporting has been in place for a number of years. Contract negotiation with the AI supplier must ensure that the software can be tested on retrospective data, or on data from outside the hospital. MPEs must be adequately trained in terms of database handling and the appropriate statistical evaluation.

Advantages: the use of local data bases gives a direct feedback on the applicability of the AI tool.

Limitations: creating an appropriate data set may require a lot of time, especially in the absence of structured reporting. It may not be easy to get access to rare or subtle cases. Gold standard results may not be available. Local or independent radiologists may have to define the truth. Moreover, databases might need to be created for every AI application that has a different use.

Example: local applications of particular software tools are sometimes reported at conferences. Typical examples include the automatic calculation of mammographic density evaluation. These studies are possible due to structured reporting in mammography, the specific age ranges and the control of image quality that is common practice in breast cancer screening. The study by Shah *et al* [28] describes such an application.

Simulate a set of test data as in virtual clinical trials (so far mainly applied to software used for supporting diagnosis, lesion detection or organ segmentation)

A third alternative is gaining in importance since the publications by Sharma and Badano [29,30]: the FDA supports the development and use of technique referred to as 'in silico testing' or 'virtual clinical trials' (VCTs) to judge the quality of (new) x-ray devices. Virtual clinical trials are consistent with the recent evolution towards 'task-based testing' and have recently been reviewed [31]. Application of this method requires: selection of the most difficult task for the AI application, to derive test signals that represent the particular situation and to create simulated images with these critical signals. In VCTs, critical signals cover a range of strengths and for different backgrounds. These signals can act as the de-facto gold standard. The testing can focus on local quality levels and local requirements.

Advantages: very detailed results can be obtained, subdivided for different lesion types that are all very well characterized. As soon as the signals have been modelled and if a few normal backgrounds or realistic virtual models of patients are available, image creation can start. It is then easy to generate large data sets. There are no extra patient exposures involved.

Limitations: the development of a simulation platform can be time consuming and requires the help of the x-ray device manufacturer to generate routine output images of these test images. It must also be verified whether the images are sufficiently realistic and whether AI tools perform satisfactorily when applied to simulated images.

Example: the approach to VCT for studies in digital mammography was initiated as far back as 2003 [32]. A further example is the performance assessment of different processing algorithms in 2D digital mammography and where it has been shown that image processing may impact the detection performance of microcalcifications [33]. The same images may be useful for testing AI based solutions. In a recent paper, an AI algorithm has been applied to detect simulated Covid-19 pathology in simulated x-ray images [34]. Simulated images have been used to train AI algorithms for detecting the COVID-19 from 2D chest x-rays [35] and it is likely that this approach will be implemented for many more applications in rare diseases or for emerging technologies, making it a topic of interest for the MPE. Sorin et al [36] review the use of Generative Adversarial Networks to train other algorithms, but potentially also useful to generate artificial images to test AI algorithms.

Commissioning tests require extensive efforts on behalf of the team, in which radiologists, medical physicists and data scientists should all ideally cooperate. Shah et al [28] state that "radiologists may wish to focus validation on special cases where expertise is challenged or has faltered (i.e. cases with subtle findings or known missed diagnoses – clinical value-added scenarios); in contrast, data scientists prefer validation using balanced and representative case cohorts instead." Medical physicists have a unique role to play if simulated images are required to enrich test databases, as discussed in subsection (3).

Once the commissioning is performed, the further roll out the AI tool in the clinical workflow requires a communication plan and adequate training for the end users. Training of the stakeholders is key for the software adoption and should address how to use expected results and how the AI solution may impact the daily work. It should also be specified whether algorithm output can be overruled or amended, as a means of improve the model itself. Feedback opportunity should be foreseen in a structured approach, including possible incidents.

3.4. Quality assurance of AI solutions

AI tools that are applied to radiological images to improve their diagnostic or therapeutic results, or to reduce radiological workload,

should also be included in a QA scheme, after commissioning. Results achieved over several months may make up the report to the management board and act as direct motivation for further investment. For the wider community, scientific reports on the long-term use of the AI solutions may be very interesting.

It is recommended that suitable KPIs are defined, so that impact on either practical issues, as well as patient outcome or any other quality parameter, can be shown. Typical practical examples include the number of cases upon which the AI tool could be successfully applied (and what was the cause of failures), user friendliness of the tool, uptime and downtime and impact on work load. The KPIs of highest importance try to measure the impact on patient care. The user could put a hard limit on a KPI or use the KPI to compare between the different solutions.

If publication of quality data is an aim, even when data are taken from routine applications, it is important to discuss this with the ethical committee upfront. Several (local) ethical rules may prevent an analysis from being published and important information for other healthcare providers would be lost.

IAEA textbook [37] proposes a scheme for Quality Management Systems (QMS) that could serve as an example of a structure for the QA of AI software tools. We suggest the following adaption to the basic steps that has been established for x-ray devices:

(i) Identify the processes needed to ensure quality of the AI application throughout the organization.
(ii) Determine the criteria and methods needed to ensure that both operation and control of these processes is effective.
(iii) Ensure the availability of resources and the information necessary to support operation and monitoring of these processes.
(iv) Monitor, measure and analyze the results.
(v) Implement actions necessary to achieve planned achievements and continual improvements of the processes.

There are many concrete reasons to implement a QA program for software tools. Some AI tools might be subject to what is called external drifting of the model: the data input may change over time (e.g. new scanner, changes in population mixture) and therefore tests performed during acceptance or commissioning may be obsolete. Models developed on limited sample sizes may initially incorporate some systematic bias that will be gradually reduced over time as the models are fed with progressively larger datasets for training and validation.

Another possible outcome can be internal drifting of the model, where the AI model itself changes over time. While some models are 'locked', e.g. they remain static over time and always provide the same output when feed with the same input, others might be adjusted by user who may select different working points, e.g. to balance sensitivity vs. specificity of an algorithm. The AI algorithm might learn in the field such that an updated model can be activated via explicit manufacturer or user interaction.

These are key points for the MPE when setting in place a QA program for such tools as these changes are often impossible to forecast during acceptance testing and commissioning, and represent potential source outcome inconsistencies of the software. As far as we know, the FDA has not yet approved any continuously learning algorithm for medical applications. The AI application may need to be tested and certified annually, as is done for physical medical-radiological devices by the medical physicist.

It is also important to note that, if a natural or legal person (1) changes the intended purpose or (2) modifies a device already placed on the market or put into service in such a way that compliance with the applicable requirements is affected, then that person must assume the obligations incumbent on manufacturers. In other words, when introducing a commercially available AI tool in a clinical workflow, the application must be used as intended, and so this assessment should also be part of the QA of the system.

Please note that the proposed concept is applicable to any software solution impacting diagnosis or treatment, not only to AI based software. Quality managers in larger hospitals can take part in this particular QA activity. In practice, however, medical image analysis is usually not part of their routine activities, while it is an important part of medical physics. Image analysis competences may be needed to either obtain the QA score, interpret the data or for trouble shooting.

4. Discussion

Artificial Intelligence based solutions can improve patient outcomes; this should be the ultimate result of the revolution currently occurring in medical imaging. Some of the new AI solutions will soon be established in routine practice. Well known candidates are software packages that search for specific lesions in specific types of images, such as acquired in the frame of organized screening [38,39] but many questions must be answered before they gain wide spread use in organized breast cancer screening [40] and in other domains of medicine [41]. Other AI tools also have the potential to achieve global acceptance, in many domains of medicine [42–44]. Typical results are lower noise (reconstructed) images, segmentation of specific anatomical parts of interest, quantitative data of all (cancerous) lesions rather than subjective readings of only a selection, quantitative lesion follow up, new data or insights in new diseases, the comparison of quantitative image data to that of other patients or to asymptomatic groups to better situate the patient's disease or prognosis, and improved outcome in general. In a recent paper, Oren et al. point the need to address clinically meaningful outcomes [45]. Other reasons for procurement of AI tools, and possibly equally determining, are aspects such as improved or faster workflow, greater consistency, faster or less tedious ways to a diagnosis or treatment plan, repeated and time consuming measurements taken over by an automated algorithm, preparation of a (structured) report.

Historically, MPEs have typically been involved with strictly technical studies [8–12] but have recently expanded their remit to include also AI methods. Therefore it could be argued that MPEs may not be well placed to interfere in the acceptance of tools that have a more direct clinical endpoint such as clinical decision support systems. However, MPEs have successfully helped to validate software tools in the past that have clinical outcome. In addition, the results of many 'technical' studies are increasingly expressed in terms of lesion detectability or any other clinical tasks, in accordance also with the Medphys 3.0 philosophy [22,23]. The following references are illustrative for relevant clinical endpoints to earlier technical studies: [33,46–48]. In the paper by G. Mahadevaiah et al. [49], a systematic approach to the different aspects involved in the adoption of decision support systems is given. There are many similarities with our approach. A list of actions is given, with many of the steps technical in nature, on the bridge between medicine and technology. It is recognized that a proper QA plan "will help avoid pitfalls, improve patient safety, and increase the chances of success". While the role of checking the clinical performance of a decision support system is ultimately with the radiologist, this will only be effective if supported on the technical and statistical front by a wider team. This applies to many more applications of AI algorithms than just decision support programs. Blazis et al [50] provide an example of an important study by an MPE, preparing the local use of an AI tool: the authors compare the impact of image reconstruction quality on the performance of a chest CT reader.

Regulatory agencies and professional societies will likely be very important stakeholders in the process of assessing AI tools. There are several initiatives ongoing like the ACR working closely with the FDA "to ensure that certification standards are in line with their approval criteria to minimize the time involved in the regulatory process". They list all FDA approved AI algorithms [https://models.acrdsi.org]. Note that unfortunately and as far as we know, a database of AI algorithms concluded to be in agreement with the MED [4] and the MDR [16] does not yet exist. As concluded in [51], evidence on the safety and performance of the European approach to the regulation of medical devices,

including AI/ML-based medical devices, is scarce; the information submitted to Notified Bodies and regulators to get CE marking is confidential. Also, the CE label does not guarantee conformity with the MED and does not ensure optimized patient care. As far as we know, a database of AI algorithms concluded to be in agreement with the MED does not yet exist and there is no test protocol. The knowledge of regulatory requirements beyond radiation safety should be part of the (future) core competences of the MPE. A whole new curriculum is being established [52]. The interested reader can refer to a recently published article on the subject [53].

Given the multitude of possible applications, no single software solution can address all tasks and therefore AI solutions are very task specific, often with a narrow domain of application and with testing limited to specific subgroups of patients. One of the associated challenges is that many software packages may have to be integrated in the IT backbone of a hospital, from different companies, both large and small. There is a risk of fragmentation, with every new AI tool requiring a very different approach at installation and requiring repeated, substantial (medical physics) QA efforts. This fragmentation could potentially lead to suboptimal governance of the software. In practice, getting started with AI can be hampered at many stages. PACS managers may oppose multi-vendor evolution and impose harmonized packages. Other challenges include (lengthy) (local) processes to get access to data for test purposes, and conversely, negotiations between the legal department and the company if the data should prove valuable to the company in refining the AI tool. General quality management approaches may be very welcome at this stage. MPEs are encouraged to expand their core activity of dose and image quality monitoring to include the new AI tools [3,4]. The successful integration of promising AI tools in processes of care can make the MPE a partner of the management board that is these days confronted with important steering decisions with regard to this digital revolution. Ultimately, the MPE may have to be prepared to reach out to other sources of information (data) in the hospital, and guide further developments towards data linking between, e.g. imaging and genomics. This was recognized by Martín-Noguerol et al. [54], who gave 'software engineers and data scientists' a central role. An extra dimension can also come from requests of network hospitals investigating similar AI tools.

MPEs should help a hospital choose AI solutions that have the greatest potential for added value for specific groups of patients. Given the risk that an inaccurate or inappropriate tool deteriorates the quality of healthcare and puts patients at risk, then safeguards are required. Stringent procurement, acceptance testing, commissioning and quality assurance must be put in place. This task is however not easy and not yet well developed. MPEs must be trained with regard to basic and advanced concepts of AI algorithms and applications [52,55] and have to team up with the medical teams for performing these tasks. It is also extremely important that regulatory agencies and industry jointly develop robust AI standards of practice and transparent insertion rules [56]. MPEs may be able to develop their test methods from the actual management procedures applied to x-ray devices. Hence, the ultimate goals are similar: ensure consistent quality. The number of scientific papers describing how to approve software solutions applied to imaging is currently very limited. In this paper, we have listed several aspects to be taken into account, along with some basic approaches for developing test methods. MPEs should be able to verify the quality of the input data or guide any double-blind trial studies or, if judged necessary, launch task specific virtual clinical trials. Given the effort needed to run such studies, the community would be helped with some type of accreditation, sharing of experience and standardization. Virtual clinical trial testing requires also the help of manufacturers in processing or reconstructing simulated data. Two aspects would significantly speed up these methods: (1) A larger EU consortium, a project or institution to collect and validate these outputs while repeated work is avoided, and (2) specific teaching modules for the different players at the introduction of these tools, with specific modules addressing the MPEs. In this text, we have tried to develop a framework for the tasks the MPE community is facing.

Finally, 'Garbage in – Garbage out' (Gigo) is probably the best known and most feared problem of AI development and implementation. In medical imaging, the quality of AI solutions relies on the technical and radiographic quality of the input images. Therefore the MPE's quality management systems of x-ray imaging data on the one hand and the AI tools on the other hand may have to be linked. As an example, there may be restrictions on altered dose settings or the adjustment of CT reconstruction settings if the images are fed directly to AI tools. The MPE must be aware of image quality assessment and the means to monitor image quality.

A limitation of this paper is the lack of an established protocol or the inclusion of concrete examples. This highlights the need for more work and for MPEs to share experiences. These days, commercially available tools usually solve concrete applications, but this is not the end-point. Automated data retrieval opens up a large number of applications, way beyond the focused outputs currently supported: data can be used in big data analytics and data mining, possibly combined with other '-iomics' (such as genomics). This may allow the detection or prediction of features or pathologies not envisaged at the design stage, leading to the next breakthrough in medicine. The MPE has a role to play in the appropriate organization and storage of these data.

5. Conclusion

MPEs can make a very relevant and important contribution to the integration of AI tools in the radiology department. In fact, in some countries their involvement is a legal requirement. Hence, MPEs are well placed to form a bridge between manufacturer and medical team, from procurement all the way through to reporting selected KPIs to the management board. One of the main challenges at this point in time lies in the development of an acceptance and commissioning test framework that can cope with the enormous range of AI applications available. With the right protocols, the MPE can help reduce the lag between the medical question and a fully integrated AI application providing the answer. Of equal importance is that the MPE sees beyond the current, focused use of AI applications in imaging, to the broader picture of AI in medicine.

Acknowledgement

We acknowledge the help of prof. N. Marshall for improving the text of this paper.

References

[1] Tang X. The role of artificial intelligence in medical imaging research. BJR Open 2019;2(1):20190031. https://doi.org/10.1259/bjro.20190031.
[2] Brink J, Arenson R, Grist T, Lewin J, Enzmann D. Bits and bytes: the future of radiology lies in informatics and information technology. Eur Radiol 2017;27(9): 3647–51. https://doi.org/10.1007/s00330-016-4688-5.
[3] Caruana CJ, Tsapaki V, Damilakis J, Brambilla M, Martín GM, Dimov A, et al. EFOMP policy statement 16: the role and competences of medical physicists and medical physics experts under 2013/59/EURATOM. Phys Med 2018;48:162–8. https://doi.org/10.1016/j.ejmp.2018.03.001.
[4] Council of the European Union. (2013). Council Directive 2013/59/Euratom laying down basic safety standards for protection against the dangers arising from exposure to ionising radiation, and repealing Directives 89/618/Euratom, 90/641/Euratom, 96/29/Euratom, 97/43/Euratom and 2003/122/Euratom. Official Journal L-13 of 17.01.2014.
[5] BELGISCH STAATSBLAD/MONITEUR BELGE 20.02.2020, p 10094 – 10154. Koninklijk besluit betreffende de medische blootstellingen en blootstellingen bij niet-medische beeldvorming met medisch-radiologische uitrustingen van 13 FEBRUARI 2020 / Arrêté royal relatif aux expositions médicales et aux expositions à des fins d'imagerie non médicale avec des équipements radiologiques médicaux de 13 FEVRIER 2020.
[6] Kortesniemi M, Tsapaki V, Trianni A, Russo P, Maas A, Källman HE, et al. The European Federation of Organisations for Medical Physics (EFOMP) White Paper: Big data and deep learning in medical imaging and in relation to medical physics profession. Phys Med 2018;56:90–3. https://doi.org/10.1016/j.ejmp.2018.11.005.

[7] The European Commission, Directorate-General for Energy Directorate D. RADIATION PROTECTION N° 162 Criteria for Acceptability of Medical Radiological Equipment used in Diagnostic Radiology, Nuclear Medicine and Radiotherapy, 2012. ISBN 978-92-79-27747-4. doi: 10.2768/22561.

[8] Singh R, Wu W, Wang G, Kalra MK. Artificial intelligence in image reconstruction: the change is here. Phys Med 2020 Nov;79:113–25. https://doi.org/10.1016/j.ejmp.2020.11.012.

[9] Wang T, Lei Y, Fu Y, Curran WJ, Liu T, Nye JA, et al. Machine learning in quantitative PET: a review of attenuation correction and low-count image reconstruction methods. Phys Med 2020;76:294–306. https://doi.org/10.1016/j.ejmp.2020.07.028.

[10] Sheng K. Artificial intelligence in radiotherapy: a technological review. Front Med 2020;14(4):431–49. https://doi.org/10.1007/s11684-020-0761-1.

[11] McCollough CH, Leng S. Use of artificial intelligence in computed tomography dose optimisation. Ann ICRP 2020;1. https://doi.org/10.1177/0146645320940827. 146645320940827.

[12] Unkelbach J, Bortfeld T, Cardenas CE, Gregoire V, Hager W, Heijmen B, et al. The role of computational methods for automating and improving clinical target volume definition. Radiother Oncol 2020;S0167–8140(20):30838. https://doi.org/10.1016/j.radonc.2020.10.002.

[13] El Naqa I, Haider M, Giger ML, Ten Haken R. Artificial Intelligence: reshaping the practice of radiological sciences in the 21st century. Br J Radiol 2020;93(1106):20190855. https://doi.org/10.1259/bjr.20190855.

[14] Nensa F, Demircioglu A, Rischpler C. Artificial intelligence in nuclear medicine. J Nucl Med 2019;60(Suppl 2):29S–37S. https://doi.org/10.2967/jnumed.118.220590.

[15] Hamamoto R, Suvarna K, Yamada M. Application of artificial intelligence technology in oncology: towards the establishment of precision medicine. Cancers (Basel) 2020;12(12):3532. https://doi.org/10.3390/cancers12123532.

[16] Official Journal of the European Union L117/1. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32017R0745&from=EN. Regulation (EU) 2017/745 of the European parliament and the council of 5 April 2017 on medical devices.

[17] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27. on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC. General Data Protection Regulation, GDPR; April 2016.

[18] Metz CE. Receiver operating characteristic analysis: a tool for the quantitative evaluation of observer performance and imaging systems. J Am Coll Radiol 2006;3:413–22. https://doi.org/10.1016/j.jacr.2006.02.021.

[19] Rogers EM. Diffusion of innovations. New York: Free Press; 1983.

[20] Dawid H, Decker R, Hermann T, Jahnke H, Klat W, König R, et al. Management science in the era of smart consumer products: challenges and research perspectives. CEJOR 2017;25:203–30. https://doi.org/10.1007/s10100-016-0436-9.

[21] De Rosis S, Nuti S. Public strategies for improving eHealth integration and long-term sustainability in public health care systems: findings from an Italian case study. Int J Health Plann Mgmt 2018;33:e131–52. https://doi.org/10.1002/hpm.2443.

[22] Samei E, Pfeiffer DE. Clinical imaging physics: current and emerging practice. 1st ed. John Wiley & Sons; 2020.

[23] Samei E. Medical Physics 3.0: Ensuring Quality and Safety in Medical Imaging Health Phys 2019;116(2):247–55. doi: 10.1097/HP.0000000000001022.

[24] Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet 2019;393:1577–9. https://doi.org/10.1016/s0140-6736(19) 30037-6.

[25] CONSORT-AI and SPIRIT-AI Steering Group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. Nat Med 2019;25:1467–8. https://doi.org/10.1038/s41591-019-0603-3.

[26] Liu X, Faes L, Calvert MJ, Denniston AK. CONSORT/SPIRIT-AI Extension Group. Extension of the CONSORT and SPIRIT statements. Lancet 2019;394:1225. https://doi.org/10.1016/S0140- 6736(19)31819-7.

[27] Bejnordi B, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017;318(22):2199–210. https://doi.org/10.1001/jama.2017.14585.

[28] Shah C, Kohlmyer S, Hunter K, Jones S, Chen P-H. A translational clinical assessment workflow for the validation of external artificial intelligence models. Proc. SPIE 11601, Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications,116010F (15 February 2021) doi: 10.1117/12.2581771.

[29] Sharma D, Graff CG, Badal A, Zeng R, Sawant P, Sengupta A, et al. Technical note: in silico imaging tools from the VICTRE clinical trial. Med Phys 2019 Sep;46(9):3924–8. https://doi.org/10.1002/mp.13674.

[30] Badano A, Graff C, Badal A, Sharma D, Zeng R, Samuelson F, et al. Evaluation of digital breast tomosynthesis as replacement of full-field digital mammography using an in silico imaging trial. JAMA Netw Open 2018;1(7):e185474. https://doi.org/10.1001/jamanetworkopen.2018.5474.

[31] Abadi E, Segars W, Tsui B, Kinahan P, Bottenus N, Frangi A, et al. Virtual clinical trials in medical imaging: a review. J Med Imag 2020;7(4):042805. https://doi.org/10.1117/1.JMI.7.4.042805.

[32] Carton AK, Bosmans H, Van Ongeval C, Souverijns G, Rogge F, Van Steen A, et al. Development and validation of a simulation procedure to study the visibility of micro calcifications in digital mammograms. Med Phys 2003;30(8):2234–40. https://doi.org/10.1118/1.1591193.

[33] Zanca F, Jacobs J, Van Ongeval C, Claus F, Celis V, Geniets C, et al. Evaluation of clinical image processing algorithms used in digital mammography. Med Phys 2009;36(3):765–75. https://doi.org/10.1118/1.3077121.

[34] Rodríguez Pérez S, Coolen J, Marshall N, Cockmartin L, Biebaû C, Desmet J, De Wever W, Struelens L, Bosmans H. Methodology to create 3D models of Covid-19 pathologies for Virtual Clinical Trials. Accepted for publication in JMI Dec. 11, 2020; published online Jan. 4, 2021. DOI: 10.1117/1.JMI.8.S1.013501.

[35] Rasheed J, Hameed A, Djeddi C, Jamil A, Al-Turjman F. A machine learning-based framework for diagnosis of COVID-19 from chest X-ray images. Interdiscip Sci Comput Life Sci 2021. https://doi.org/10.1007/s12539-020-00403-6.

[36] Sorin V, Barash Y, Konen E, Klang E. Creating artificial images for radiology applications using generative adversarial networks (GANs) – a systematic review. Acad Radiol 2020;27:1175–85. https://doi.org/10.1016/j.acra.2019.12.024.

[37] nce D, Christofides S, Maidment A, McLean I, Ng K. Diagnostic Radiology Physics: A Handbook for Teachers and Students, 2014. Downloadable from the IAEA website (dec 2020) at: https://www-pub.iaea.org/mtcd/publications/pdf/pub1564webnew-74666420.pdf.

[38] Lång K, Dustler M, Dahlblom V, Åkesson A, Andersson I, Zackrisson S. Identifying normal mammograms in a large screening population using artificial intelligence. Eur Radiol 2020. https://doi.org/10.1007/s00330-020-07165-1.

[39] Schaffter T. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. JAMA Netw Open 2020;3(3):e200265. https://doi.org/10.1001/jamanetworkopen.2020.0265.

[40] Sechopoulos I, Mann R. Stand-alone artificial intelligence – the future of breast cancer screening? Breast 2020;49:254–60. https://doi.org/10.1016/j.breast.2019.12.014. Epub 2020 Jan 2.

[41] Miller Dd, Brown E. Artificial Intelligence in Medical Practice: The Question to the Answer? Am J Med 2018 Feb;131(2):129-133. doi: 10.1016/j.amjmed.2017.10.035. Epub 2017 Nov 7.

[42] Daldrup H. Artificial intelligence applications for pediatric oncology imaging. Pediatr Radiol 2019;49(11):1384–90. https://doi.org/10.1007/s00247-019-04360-1. Epub 2019 Oct 16.

[43] Panayides A, Amini A, Filipovic N, Sharma A, Tsaftaris S, Young A, et al. AI in medical imaging informatics: current challenges and future directions. IEEE J Biomed Health Inform 2020;24(7):1837–57. https://doi.org/10.1109/JBHI.2020.2991043.

[44] Tandon Y, Bartholmai B, Koo C. Review article on contemporary practice in thoracic neoplasm diagnosis, evaluation and treatment. Putting artificial intelligence (AI) on the spot: machine learning evaluation of pulmonary nodules. JTD 2020;12(11). https://doi.org/10.21037/jtd-2019-cptn-03.

[45] Oren O, Gersh BJ. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. Lancet Digit Health 2020;2(9):e486–8. https://doi.org/10.1016/S2589-7500(20)30160-6.

[46] Warren LM, Given-Wilson RM, Wallis MG, Cooke J, Halling-Brown MD, Mackenzie A, et al. The effect of image processing on the detection of cancers in digital mammography. AJR Am J Roentgenol 2014;203(2):387–93. https://doi.org/10.2214/AJR.13.11812.

[47] Michielsen K, Nuyts J, Cockmartin L, Marshall N, Bosmans H. Design of a model observer to evaluate calcification detectability in breast tomosynthesis and application to smoothing prior optimization. Med Phys 2016;43(12):6577. https://doi.org/10.1118/1.4967268.

[48] Samei E, et al. Performance evaluation of computed tomography systems: Summary of AAPM Task Group 233. Med Phys 2019;46(11):e735–56. https://doi.org/10.1002/mp.13763. Epub 2019 Sep 11.

[49] Mahadevaiah G, Rv P, Bermejo I, Jaffray D, Dekker A, Wee L. Artificial intelligence-based clinical decision support in modern medical physics: selection, acceptance, commissioning, and quality assurance. Med Phys 2020;47(5):e228–35. https://doi.org/10.1002/mp.13562.

[50] Stephan Blazis P, Dickerscheid Dennis BM, Linsen Philip VM, Jarnalo Martins, Carine O. Effect of CT reconstruction settings on the performance of a deep learning based lung nodule CAD system. Eur J Radiol 2021;136:109526. https://doi.org/10.1016/j.ejrad.2021.109526.

[51] Muehlematter U, Daniore P, Vokinger K. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. Lancet Digit Health 2021;3(3):e195–203. https://doi.org/10.1016/S2589-7500(20)30292-2.

[52] Zanca F, Hernandez-Giron I, Avanzo M, Guidi G, Crijns W, Diaz O, Kagadis O, Rampado O, Lønne P, Ken S, Colgan N, Zaidi G, Kortesniemi M. Expanding the Medical Physicist Curricular and Professional Programme to include Artificial Intelligence. In press in Physica Medica.

[53] Beckers R, Kwade Z, Zanca F. The EU medical device regulation: Implications for artificial intelligence-based medical device software in medical physics. Accepted for publication in Physica Medica. In Press.

[54] Martín-Noguerol T, Paulano-Godino F, López-Ortega R, Górriz J, Riascos R, Luna A. Artificial intelligence in radiology: relevance of collaborative work between radiologists and engineers for building a multidisciplinary team. Clin Radiol 2020;(20):30602–4. https://doi.org/10.1016/j.crad.2020.11.113. S0009–9260.

[55] Diaz O, Guidi G, Ivashchenko O, Colgan N, Zanca F. Artificial intelligence in the medical physics community: an international survey; Accepted for publication in Physica Medica.

[56] Miller D. Machine intelligence in cardiovascular medicine. Cardiol Rev 2020;28(2):53–64. https://doi.org/10.1097/CRD.0000000000000294.