Evaluating Pre-trained 3rd Party AI Models in Medical Device Software Development for Responsible and Ethical Use

St John Lynch, N[1[0009-0009-4150-4970]], McHugh, M[1[0000-0003-4275-3302]], Loughran, R[1[0000-0002-0974-7106]], McCaffrey, F[1[0000-0002-0839-8362]], Kowalski, D[2[0009-0001-0663-8609].

(1) Regulated Software Research Centre, Dundalk Institute of Technology, Dublin Road, Dundalk, Co. Louth, A91 K584 Ireland

(2) Brainlab SE, Olof-Palme-Straβe 9, 81829 Munich, Germany

niamh.stjohnlynch@dkit.ie

Abstract. It is well understood that new standards are necessary to ensure the use of Artificial Intelligence-enabled Medical Devices (AIeMD) are adequately controlled. CEN-CENELEC (European focus) and IEC (Global focus) are developing such standards. Yet the use of pre-trained AI models, available as Software of Unknown Provenance (SOUP) from 3rd party suppliers, remain excluded from the standards under development. This paper aims to demonstrate the need for standardisation of the qualification process for these models beyond that understood by existing standards such as ISO 13485 and IEC 62304. This is necessary in order to achieve responsible and ethical AI for use in healthcare. This paper demonstrates the need for guidance by setting out emerging use cases as well as the risks and potential risk mitigations available when using pre-trained models as SOUP in the Software Development Lifecycle of AIeMD. It goes further to outline the necessary characteristics that should be standardised. Evaluation questions are provided and demonstrate salient points necessary for responsible and ethical AI.

Keywords: Pre-trained Artificial Intelligence, transformer, foundation model, SOUP, Software Development Lifecycle, SDLC, Medical Device.

1 Introduction

1.1 Pre-trained AI Models use in Medical Devices

Pre-trained Artificial Intelligence (AI) models are becoming the go-to solution for a range of AI challenges in healthcare. A pre-trained AI model is an AI foundation model that has already been trained on a large dataset to learn general patterns of language, image or audio or can be a combination of these. They can be used as is, or adapted for specific tasks. Pre-trained models are becoming more widely available and save an AI developer time and resources. General Pre-trained Transformer (GPT)[1] and Bidirectional Encoder Representations from Transformers (BERT)[2] are examples of Large Language Models (LLMs) that have been trained on large open-source data sets [3]. The release of BERT by Google in 2018 was the first encoder architecture to read a sentence from both directions [2]. This was a significant improvement over previous models, including GPT at the time, that read text from left to right. BERT was deemed to be able to better understand the context and therefore, be a better predictor of text responses. BERT is trained on a huge dataset (e.g., Wikipedia) and can be subsequently fine-tuned for specific tasks with relatively little data. Although BERT was a game-changer, when we consider its use in AI-enabled Medical Devices (AIeMDs), we need to be cautious. As suggested by the Google team who first presented BERT, fine-tuning can be done by adding just one additional output layer to create state-of-the-art models for a wide range of tasks [2]. Any pre-trained model (PTM) can be expected to have in-built bias given the wide range of data used, including 'knowledge' that may be inaccurate. Although BERT is a significant improvement on previous models, when used for development of medical devices, there could be any amount of misinformation in the pre-learned dataset, such as "*alkaline diet is a cure for cancer*" [4]. This requires a level of rigor and independent assessment by competent personnel to evaluate the model selected and assess potential for in-built bias in the pre-trained AI model. It also raises questions as to how much additional data and training should be used in the transfer and fine-tuning process to ensure the AIeMD is adequately trained to perform as intended in healthcare [5], [6]. Lastly it raises questions about the level of transparency provided by 3rd party suppliers and what information should be passed onto industry and eventually regulators when reviewing technical documentation for market clearance to meet regulatory requirements [7], [8].

This paper focuses on PTM use by medical device organizations and AI developers seeking to reap the many benefits promised from AI in healthcare [9]. The aim is to inform the healthcare industry, standards organizations and regulatory bodies about the advantages and disadvantages of using PTMs. It highlights the risks that can arise when selecting and using PTMs [10]. It also recommends a standardised approach for evaluating these models to ensure responsible, ethical use with safety and security. To support standardisation, this paper provides a guidance checklist that can be used to evaluate PTMs when adopted for use in AIeMDs. This checklist reflects on current research such as Trustworthy AIeMD including questions adopted from the AI-Change Management (AICM) process [11]. The AICM checklist includes a section on PTMs and this is reviewed against the Chinese standard issued by the NMPA, China's Competent Authority [12].

Europe and the US have yet to create such a standard. In the meantime, development of an AIeMD using PTMs is well underway. This is recognized by the Consumer Technology Association (CTA), who have written to the FDA to advise on the need for consideration of PTMs following the recent FDA guidance "*Artificial Intelligence Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations*" and "*Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products*" [13]. The CTA have recommended that the FDA clarify the regulatory obligations for 3rd-party developers of foundation models (i.e., PTMs) used in AIeMD, emphasizing that these developers should *not* be subject to Quality System Regulation requirements. The CTA urge a risk-based approach. Where the responsibility lies on the medical device manufacturer, the requirements of ISO 13485 Quality Management System (QMS) are not sufficient without additional clarity and regulatory guidance for mitigating the risks posed by pre-trained AI models [14].

The objective of this paper is to outline the prevalence of PTM use in medical device development. It seeks to highlight the benefits and risks associated with selection and use and provide a method for standardising the assessment as part of the Software Development Lifecycle (SDLC) process. This paper is broken down as follows: ***Section 2*** identifies example use cases in healthcare using pre-trained AI models. ***Section 3*** demonstrates the variety and accessibility of pre-trained AI models available to medical device developers. ***Section 4*** looks at pros and cons with a deeper dive into risks associated with PTMs. ***Section 5*** looks at current guidance and standardisation for these models in the medical device SDLC. ***Section 6*** examines some of the primary characteristics that require guidance for understanding when using these models. The conclusion is presented in ***Section 7*** followed by an ***Appendix***, with a list of considerations to be utilized by a regulator and AIeMD developer when selecting and documenting a PTM for use in the SDLC.

2 Pre-Trained Model Use Cases in Healthcare

2.1 Pre-trained Large Language Models

Research demonstrates that PTMs can be used with limited additional training to solve a task 'similar' to that in which it is trained typically requiring 'fine-tuning'. For use in healthcare, the question of the amount of additional learning and fine-tuning is important. It is understood that the model architecture can be enhanced by transformer models for encoding and decoding patterns, hence paving the way for a myriad of solutions [15]. The BERT example above is an encoding only model. The question of whether the architecture needs to be updated is a consideration for developers. Examples of uses cases include *medical note-taking* or *prescription writing* based on auditory recognition of a patient-practitioner consultation, with pre-training performed on a wide range of open-source data. The medical device manufacturer can decide to narrow the scope and fine-tune the AI model with additional medical-specific learning and testing. Guardrails should be set to ensure the LLM continues to learn from peer-reviewed clinical articles or a defined set of medically accepted text books. However, bias may already have been learned from the pre-learning stage [16]. Interestingly Ada Health, an SaMD AI symptom checker is not built on a generic pre-trained LLM like GPT or BERT [17]. Instead, it uses specialized knowledge based with proprietary reasoning algorithms that were developed with specialist medical knowledge in-house. This type of system is closer to a clinical decision support system and not considered a general-purpose LLM. The purpose of highlighting this is that approval of one type of 'diagnostic' clinical support tool may not reflect the architecture of another in any way. This type of AI is covered under the current regulatory framework and 'Ada' is cleared for market use in the EU having a product risk category of IIa (i.e., low risk device) [18]. Standards under development are focusing on introducing requirements for machine learning (ML) having a narrow focus, and only just starting to initiate guidance and standards on LLMs [19]. PTMs are yet another category in their own right. This highlights the difference in architecture that is important when assessing the risk-profile of any AIeMD.

2.2 Pre-trained Large Imaging Models

There may be an assumption that this is a problem for language models only. However, research demonstrates that the majority of use-cases in healthcare currently involve image classification, particularly radiology and cardiology. This is primarily due to the number of images already collected by both healthcare organization and medical device manufacturers ready to reap the benefits promised from AI in healthcare[11]. PTMs used for imaging include ResNet [20] and Visual Geometry Group (VGG) [21] that have already learned features to recognize such as shapes, edges and objects from large data sets [22]. ResNet is an example of a network that has been trained on a large image dataset that can be utilized as a starting point for specific medical device use cases. One can download the weights for ResNet50 for instance, and modify the final layers of the network (referred to as retraining or transfer learning) in order to tackle a new problem scenario [23]. The VGG architecture comes in various sizes (e.g., VGG16, VGG19). VGG 16 has 16-layers (13 convolutional layers and 3 fully connected layers). These deep learning models are intended to improve accuracy by extending the depth of weights as shown in 2014 by Simonyan and Zisserman when using ImageNet (a database of 14 million images across 1000 classes) [24]. The convolutional filters are stacked to allow the model to learn complex features more effectively. Medical devices focusing on radiology typically use these types of models as their starting point [25]. The question is whether they have sufficient understanding to choose the most appropriate model and adapt it to their use case or whether they rely on whatever is available. Any AI developer can select a PTM to set about development activity and transfer learning to a new use case to improve healthcare. The intention is sound and also necessary given a healthcare system that is extensively challenged [26]. The question we must ask is, how are these models controlled and do we understand them sufficient to mitigate the risks they could introduce?

2.3 Do Pre-Trained Models identify as SOUP in Industry?

Some might simply say these PTMs are simply SOUP items and therefore should be controlled as such. Since these models are available from a 3rd party supplier they are identified as 'SOUP' – *Software of Unknown Provenance* - within the medical device SDLC. However, the traditional SOUP requirements do not adequately cover the risks associated with PTMs [27]. Moreover, industry do not appear to be able to clearly identify PTMs when brought into the SDLC process. This is evidenced in this research whereby, members of a leading industry conference held in Boston USA in October 2025, on Medical Device Cybersecurity, demonstrated. A quiz was given to 44 members of the audience in attendance using a QR code link to MS Forms. The question of whether PTMs were used in medical device development was asked. The results identified 30% as using PTMs in their medical device product development lifecycle; 61% said no-they were *not* using PTMs and 9% were not sure if they were used, refer to **Fig. 1**. The 2nd question presented here from the short quiz asked whether a PTM was *a)* an AI component, *b)* an integrated software item *c)* a SOUP item under IEC 62304 and *d)* other. All questions were posed as quick multiple-choice questions. The results show that 34% reported that PTMs were a SOUP item. It is not clear from this study whether the 34% who believe PTMs are SOUP are the same 30% who have PTMs in development. In any event, 14% said a PTM was an integrated software item and 43% (the majority) said that it was an AI component; the remaining 6% were not sure and selected 'other'. Refer to **Fig. 1** Medical Device Industry Quiz on Pre-Trained Models. **Error! Reference source not found.**Although this quiz was a 5-question quick survey presented at a medical device industry conference with a focus on cybersecurity, it highlights the lack of clarity around identifying PTMs and hence, how to control them. This survey was performed by medical device technical experts and reveals a lack of understanding that exists in relation to PTMs when used as part of an AIeMD within the SDLC process.

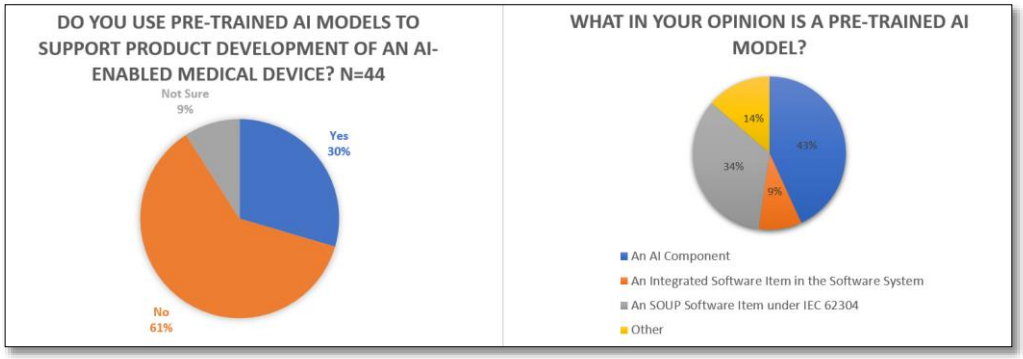

Fig. 1 Medical Device Industry Quiz on Pre-Trained Models

3 Availability of Pre-Trained AI Models

Pre-trained AI Models are widely available on the market and continue to grow given their open-source accessibility [10]. A search of Microsoft, Google or GitHub yields a wide range of PTMs for use in the medical domain. Ref. **Table 1.** LLM Medical AI/ML Pre-Trained Models for a sample.

Table 1. LLM Medical AI/ML Pre-Trained Models

Model Name	Description	Location	Use Case
BioBERT	A pre-trained biomedical language representation model.	https://github.com/dmis-lab/biobert	Biomedical text mining, NLP tasks
BioGPT	A generative pre-trained transformer model for biomedical text generation.	https://github.com/microsoft/BioGPT	Biomedical text generation, NLP
BioMistral-7B	A Collection of Open-Source Pretrained Large Language Models for Medical Domains	https://huggingface.co/BioMistral/BioMistral-7B	Biomedical text mining, NLP tasks
BlueBERT	BERT-based model trained on PubMed abstracts and MIMIC-III clinical notes.	https://github.com/ncbi-nlp/bluebert	Biomedical research, clinical notes
Clinical-BERT	A Pretrained Model on a large corpus of 1.2B words of diverse diseases. Fine-tuned on 3 million patient records.	https://huggingface.co/medicalai/ClinicalBERT	Medical question-answering (QA) tasks
Clinical-BERT	BERT model fine-tuned on clinical notes.	https://github.com/EmilyAlsentzer/clinicalBERT	Clinical text analysis, EHR data
LLaVA-Med	A large language and vision model trained using a curriculum learning method for adapting LLaVA to the biomedical domain.	https://huggingface.co/microsoft/llava-med-v1.5-mistral-7b	Medical question-answering (QA) tasks
SciBERT	A BERT-based model trained on scientific text.	https://github.com/allenai/scibert	Scientific literature analysis

3.1 Open and Closed Source Vision Models

With the growing number of open source and closed source applications emerging, widespread use by AIeMD organizations can be anticipated. **Table 2** and **Table 3** provide examples of vision and multimodal models. There are also specialised models (e.g., COVID-NET) and framework models (e.g., MONAI, PyHealth, NVIDIA Clara) among the growing number of medical models available for download.

Table 2. Pre-Trained Vision Models

Model Name	Description	Link	Use Case
Vision Transformer (Open-source, (OS))	A deep learning model for classification & segmentation on imaging data.	https://github.com/google-research/vision_transformer	Radiology, Pathology
CheXbert (OS)	Automatic Labelers and Expert Annotations.	https://github.com/stanford-mlgroup/CheXbert	Radiology
UNet (OS)	A convolutional neural network for biomedical image segmentation.	https://github.com/zhixuhao/unet	Biomedical image segmentation
nnU-Net (OS)	A self-configuring method for biomedical image segmentation.	https://github.com/MIC-DKFZ/nnUNet	Biomedical image segmentation
AI for Breast Cancer (CS)	AI model for improving breast cancer screening accuracy	https://health.google/caregivers/mammography/	Breast cancer screening, diagnostics
AI for Diabetic Retinopathy (CS)	Detects diabetic retinopathy and diabetic macular edema	https://health.google/caregivers/arda/	Diabetic retinopathy screening, ophthalmology
Sepsis Prediction Algorithm (CS)	Early warning system for sepsis in ICU patients	https://www.hopkinsmedicine.org/	Sepsis detection, ICU management
COVID-19 Severity Prediction	AI models for predicting COVID-19 severity	https://www.hopkinsmedicine.org/	COVID-19 patient management, severity prediction
InnerEye	AI for medical imaging, including radiotherapy and image analysis	https://www.microsoft.com/en-us/research/project/medical-image-analysis/	Radiotherapy planning, medical imaging analysis
PathAI	Models for pathology image analysis, including cancer detection	https://www.pathai.com/	Pathology diagnostics, cancer detection
RAD AI	AI-powered assistants for radiologists	https://www.radai.com/	Radiology assistance, diagnostics
Butterfly iQ+	Handheld ultrasound device with integrated AI for diagnostics	https://www.butterflynetwork.com/	Point-of-care diagnostics, ultrasound imaging

Table 3. Pre-Trained Multimodal Models

Model Name	Description	Link	Use Case
edNLI	Natural language inference dataset for the clinical domain.	https://github.com/jgc128/mednli	Clinical decision support, NLP
Med3D	Pre-trained 3D medical image analysis model.	https://github.com/Tencent/MedicalNet	3D medical image analysis

We also see the emergence of an Open Medical-LLM Leaderboard from 'huggingface.co' which may provide some level of assuredness, but to what extent? The leaderboard is based on reported accuracy for a given use case, at a point in time, given precise settings and objectives. The reported level of accuracy has to be cautioned, as it is difficult to re-create accuracy for different use cases, or even for the same use case. This is particularly true when a measure of accuracy of a model can be inflated based on repeat testing (overfitting) and is a measure at a point in time only and dependent on the test data. It is not always clear what is included in the accuracy measurement, e.g., whether the metric reflects the macro measure of all datasets or the best selected dataset and so on. It can only be taken as a rough guide at best and even then, must be challenged.

4 Advantages and Disadvantages of Using Pre-Trained Models

When we consider the range of medical PTMs, it starts to become clear that to ignore PTMs and start development of an AI model from scratch might be non-sensical in some cases. The advantages are the ease of access to such models ready for download. The groundwork for the AI developers is largely done, at least from a computational resource standpoint. A search for a similar intended use allows a developer to move straight to transfer learning, fine-tuning and testing with a limited dataset.

Considerations by developers that make using PTMs a valid option include improved performance where pre-training on diverse data have already encoded knowledge through pattern identification [28]. Computational resources are another key attribute that is supported. Pre-trained foundation models demand extensive compute power and trustworthy deployment must account for infrastructure and scalability. Microsoft specifically warn that their "*healthcare AI [foundation] models are intended for research and model development exploration and are not designed or intended to be deployed in clinical settings as-is*" [29]. However, using a PTM is preferred by many, as training from scratch is power intensive requiring large data stores that are difficult to source and consume high energy costs. Fine-tuning of PTMs on domain-specific data is clearly more practical requiring moderate GPU resources. While PTMs are advantageous, their adoption for clinical tasks require consideration for computing power and storage and can require high-end GPUs (e.g., NVIDIA V100/A100/H100). The release of NVIDIA's H100 allows for LLMs such as BioGPT [30] and ClinicalBERT [31] to be trained for use in healthcare tasks with faster convergence or learning (i.e., training weights stabilise and loss/error stops decreasing significantly). This enables potential for real-time imaging analysis supporting generative AI use in healthcare [32].

Residual Neural Network (ResNet) as already mentioned was introduced in 2015 for image recognition and was shown to outperform on previous visual recognition architectures such as VGGNet, released only a year before [5], [21]. ResNet is trained on large image datasets such as ImageNet (~14 million images). Patterns and feature abstraction is performed through statistical reasoning or identification of relationships (regression/correlation, etc.). Optimization and further fine-tuning are left to the AI developer. It is essential that the developer understand the coarseness of the pre-learning performed and be aware of assumptions and risks that can arise to ensure mitigation. With pre-learning optimized, large labelled datasets may not be necessary, avoiding the tedious work of data labelling required with supervised and semi-supervised AI models. This together with the performance benefits and the adaptability of these PTMs, we can see how adaptable they can be in a wide range of tasks [3]. Pre-trained AI models are therefore considered the 'ready-made' SOUP component that can provide the foundation for specific AI tasks. According to current standards, they require no more control than existing SOUP in today's medical device SDLC activities (e.g., APIs or development tools). Nevertheless, they bring with them multiple risks that are not yet well understood in the healthcare industry [33], [34]. The next section provides high-level risk categories for consideration.

4.1 Risks and Mitigation in the use of Pre-Trained Models

It is necessary to highlight some of the key risks that can be introduced from PTMs, presented in **Table 4**. Many of these risks are reflective of AI model development generally, though our focus is to ensure the developer and regulator understand the risks they are bringing into the SDLC [35]. The risks should be understood early in the development lifecycle planned for assessment with the team versed in the origins of the model, the differences and similarities of layers, settings, data, etc. The architecture, pre-training and assumptions must be made known.

Table 4. Risks and Potential Mitigations to use of Pre-Trained Models in Medical Devices

Risk	Example	Mitigation
Security & Privacy	Data poisoning, privacy leakage, model inversion, model theft, member inference, exfiltration, backdoor attack, manipulation, exposed API and credentials, tampering, insecure components, unauthorized use of model/data.	Full assessment of pre-trained model selection and purpose with security and privacy controls in place. Model versioning, continuous monitoring, zero-trust for secure deployment and maintenance.

Risk	Example	Mitigation
Bias & Inequity	Intrinsic bias (data, spatial, temporal, collection bias), extrinsic bias (task-specific, Gen-AI biases), algorithmic bias (architecture, feature selection, loss function, metrics, training methods, competency, sampling, labelling, etc).	Risk analysis and adequate controls; including understanding source of data used in pre-learning, re-sampling controls, data balancing, independence, augmentation, filtering; de-biasing, competency; standardisation, etc.
Transparency	Awareness of pre-training model details for assessment; labelling / documentation with adequate information available	Documentation availability for assessment and maintenance, taking account of guidance in appendix.
Performance	Metric suitability, adequacy; model drift assessment, over-fitting, under-fitting, generalizability.	Refinement and evaluation for model generalizability, taking account of guidance in appendix.
Technical Suitability	Architecture type, purpose, use case fit, data suitability to use case, etc.	Assessment for suitability including layers, nodes, loss function, data, sample size, representativeness of population / problem, parameters, hyperparameters, etc.

5 Standards under Development for AIeMD

A systematic literature review has provided an analysis of standards released and under development for AIeMDs [36]. This paper confirms that there is still no sign of a work program initiated for pre-trained AI models for healthcare, despite the ongoing developmental effort by standards organizations. This includes International Electrotechnical Commission (IEC), having a global perspective and European Committee for Standardisation (CEN-CENELEC) that focuses on European regulation and aims to meet the requirements of the EU AI Act [37]. CEN-CENELEC working groups are aware of the growing need for regulation and generative AI including LLM requirements that are under development as part of Joint Technical Committee JTC21. The proposed amendment of ISO/IEC 22989 which addresses concepts and terminology touches on transformer algorithms but does not explicitly deal with pre-trained AI model requirements [38] that can utilize transformer algorithms [39]. Although ISO/IEC 42001 an AI Management System is released and can provide some level of governance over such AI models, it is not specific to PTM or indeed to healthcare [40]. Ultimately, this research agrees that when using PTMs, the "*knowledge embedded in LLMs is not immediately accessible and requires careful extraction and efficient utilization to yield effective results*" [41]. This requires full understanding of the source of the PTM where possible, including a complete set of details to evaluate suitability for use. Alternatively, in the absence of available source data, a worst-case verification/validation approach should be taken by the manufacturer to mitigate risk. This research suggests a new work item be introduced under IEC TC62, JTC21 or ISO/IEC JTC 1/SC42 as applicable, taking account of the evaluation proposed here. The characteristics required for standardisation are summarised here.

6 Characteristics Required for Standardisation

Trainability. The efficacy of a PTM is not dictated by the amount of data it is trained on. There are a myriad of considerations to be taken into account when utilizing a PTM model and then become the legal manufacturer when releasing an AIeMD for use in healthcare. For example, research demonstrates that image classification transformers (encoders and decoders) based on characteristics of each pixel in an image, can be dependent on the transformer model's performance [15]. Degradation can arise as training continues. This is known as 'catastrophic forgetting'. A model's previously learned task can deteriorate as it learns new tasks [42].

Transformer-based architectures offer improvements in semantic segmentation that are promising for applications in healthcare. The choice of transformer architecture, foundation model and bias-aware evaluation must be understood by those who are developing them in healthcare. Hence, the documentation associated with the PTM should describe the overall structure and origin of the PTM and transformer utilised. The suitability of the model should be thoroughly assessed. Consideration should be given to whether the architecture in use is similar to the structure of a publicly available one; the difference should be assessed. If it is a proprietary structure, a detailed mathematical or

structural description is necessary. Understanding the precursor of pre-training is as important as the subsequent learning transfer, fine-tuning and testing that is required to deploy an AIeMD fit for purpose.

One example in healthcare is the development of TELL 2.0 [43] which assesses cognitive deficit intended to support diagnosis. A patient undergoes language-based tasks such as reading, writing, and recall tasks. TELL 2.0 was built on RoBERTa, identified as a Robustly Optimized BERT Pre-training model. Research on RoBERTa demonstrated BERT was significantly undertrained in various tasks and it was shown that the selection of hyperparameters can have a significant impact on the final results [43]. When assessing the fine-tuned TELL 2.0 application the developers need to consider domain transfer risk and the use of tools such as PySentimiento [44], which is a python toolkit for opinion mining developed multilingually. Its use in TELL 2.0 with RoBERTa has potential for linguistic mismatches in nuance, slang, accent, semantics and domain-specific terminology [45]. Misclassification of results across demographics can lead to inequitable outcomes across different populations and cultural groups. The TELL 2.0 application, having been trained in Spanish, may be subject to yet further risks where TELL 2.0 technology is to be utilised as the basis of MemoryTell 1.0. MemoryTell, a sub-set of cognitive deficits using similar tasks to identify memory deficits in English [45]. It is therefore essential that the technology on which these AI models are built are adequately understood to identify the risks and apply appropriate controls. It is not difficult to anticipate that these risks can lead to false positives/negatives and result in patient anxiety at a minimum, where cognitive deficit is being considered. This paper does not focus however, on the standards that are already covered or in development, such as AIeMD performance evaluation, risk management and AI test methods in healthcare [46]. What it does identify is the need to understand the opportunities for risk from PTMs specifically.

Robustness. It is clear that many real-world applications struggle with accessibility of large datasets, leading to the need for transfer-learning from pre-trained AI models. Some research is emerging that supports replacing deep net architectures with simple linear models where the model has been pre-trained on large language models such as Chronos and Time-MoE [47], [48]. Despite being lightweight, superior efficiency and robustness is claimed with various sampling rates and enhanced interpretability offered from the linear model [49]. Research such as this provides some confidence that a PTM can be utilized with limited fine-tuning and suggests that no manual tuning is required for subsequent deployment. Caution is advised here where this approach is adopted by developers in medical devices, where more care is needed for selection and assessment of the model. Appropriate documentation should reflect a clear understanding of the pre-training and model parameters. The documentation should describe the data augmentation methods, model weight initialization strategies, optimizers used, and the configuration of key hyperparameters during training to name a few.

Generalization. The key to a good AI model is how it measures up in use, particularly in healthcare across multiple hospitals, clinics, settings, or user groups. Research has demonstrated that even domain-specific language models such as, the Swedish clinical language model (SweDeClin-BERT), which is a modified version of ClinBERT used for clinical note taking, required further specific fine-tuning and additional refinement and evaluation in order to achieve adequate performance measures when adopted [28]. Swedish researchers downloaded SweDeClin-BERT, as having been training for medical use, for use as a means of detecting Adverse Drug Events from patient records. Only after much refinement did they achieve performance scores in the range of 0.8 F_1 based on fine-tuning and additional annotating [28]. It is clear that any medical device manufacturer and AI developer will perform performance testing and standards are in development for this [50]. However, again no consideration is given to PTMs and the additional evaluation that may be needed in each case. It is not surprising that fine-tuning is required dependent on the use case, whereby different hospitals will document records differently; understanding the source data could help identify development challenges earlier in the lifecycle and ensure there are less challenges to overcome in the SDLC. Transparency of the PTM is necessary to understand what is to be expected in development and maintenance.

Transparency. Given the brief summary of the risks identified above, the need for transparency is highlighted and also required under the EU AI Act [37]. The claim that interpretability is improved by linear model development may be so, but do we fully understand the PTM architecture to be able to make appropriate judgements and perform adequate testing [49]. Whilst full transparency and

interpretability of the PTM architecture may not be achievable, interpretability of the clinical decision-making and diagnostic parameters must be made clear when transferred into an AIeMD. Open-source models in particular are accessible to interference of source code, algorithms, and data. Transparency is both necessary and helpful in understanding and predicting the risks and vulnerabilities. Red-teaming is recommended and is performed by publishing teams such as Google, OpenAI, Meta, etc. with the aim of disclosing vulnerabilities [51]. This activity should be assessed as part of the model selection process where possible and key details reviewed for selection and suitability.

Security & Privacy. Both model inversion attacks and membership inference attacks are of particular concern to privacy in these models [52]. The identity of sensitive data can be realized from the training data where the synthesized model is widely available. Data reconstruction can be achieved from latent vectors used in the pre-training process. These risks are known amongst specialists but may not be widely understood by users, regulators and standards bodies creating AI standards for healthcare or medical device manufacturers. GitHub and Hugging Face among others are calling for more open-source support from the EU AI Act [52]. Hugging Face openly supports the use of Model Cards to help verify the source of these transformer models. Model cards have appeared in the appendix of the recent FDA guidance on AI[53], though use of the Model Card is not widely understood or mandated by EU standards to date [54]. If not Model Cards, then what? [54] We need standardisation to ensure models adopted are secure and adequately validated, or else we need guidance and adequate controls to protect the AIeMD developer against these risks. Data poisoning and backdoor attacks are typical security risks that can be expected from open-source models. Due to the size of the dataset used in training, it is not possible to expect human intervention or inspection of all data and hence, there is a considerable reliance on the integrity of the code and data used. Research has shown us that adding a small amount of manipulated data can significantly change the model's behavior, which has become known as data poisoning [52]. Although 100% check is not generally possible, a requirement for data poisoning checks is necessary. Use of PTM expands the attack surface for an AIeMD and this requires additional controls [55].

PTM Guidance Checklist. Finally, in a bid to support AIeMD development and protect industry and healthcare stakeholders, a PTM guidance checklist is provided in the Appendix 1. There may be overlap between the existing AI development activities (e.g., performance evaluation, threat modelling, etc.) which are part of the current regulatory requirements. The checklist is provided to identify whether additional controls are warranted, depending on the source and architecture of a PTM selected for a particular use and risk case. By providing guidance in this manner, the AIeMD developer is forewarned and can incorporate necessary controls as appropriate into their QMS. This aims to reduce the overall burden for developers by integrating these guidelines into the existing SDLC and AI Development lifecycle as appropriate, as shown in **Fig. *2***.

7 Conclusion

As AI systems continue to scale in complexity and capability, PTMs have emerged as foundational assets in accelerating innovation. Research demonstrates how PTMs are helping to overcome some of the key challenges including data and processing power accessibility and efficiencies. The uptake of these PTMs cannot be underestimated. The need for thorough evaluation in development and deployment by medical device manufacturers must be further explored and standardised. Assessments must include suitability, transparency, robustness, trainability, security and generalizability. Appropriate documentation should be made available by the providers of PTMs where possible. This makes it easier for the developer to be able to adequately assess the source and content of the PTM for suitability to the intended purpose and subsequent placement on the market. In the event that 3rd party suppliers do not provide adequate information, the onus is placed on the legal manufacturer to ensure they adequately assess the risks posed from these SOUP items.

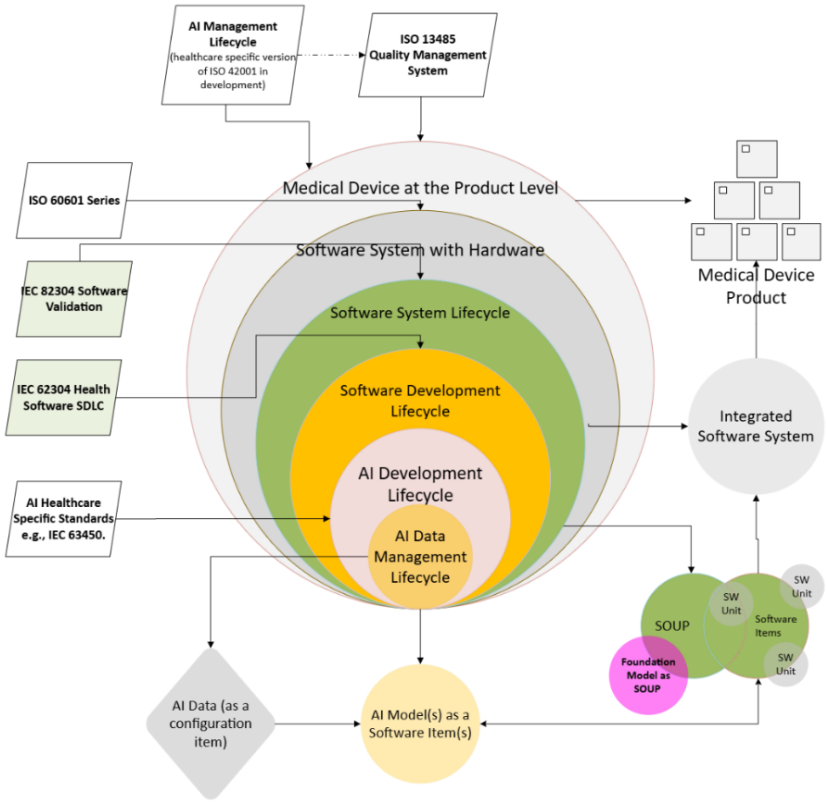

Fig. 2 Pre-Trained Model as a SOUP Item within a Medical Device SDLC

It is clear from this research that additional considerations are necessary when we define PTMs as SOUP according to IEC 62304 [27]. These are not simply traditional SOUP items, but a special kind of SOUP, that requires due consideration and additional guidance for industry. It would be highly beneficial to industry to release the checklist provided here as a Technical Report (e.g., TR by IEC TC62), which is a level below a technical specification or standard required for compliance and would therefore be considered as guidance. Whilst the intention is not to further burden industry with additional regulatory requirements, it must support industry with the requisite knowledge for protection and safety of their organisations and their customers. A collective beneficial approach to PTMs is necessary in order to build on existing quality and reap the many health benefits promised from AI, whilst delivering trustworthy and ethical AIeMDs.

Limitations. This research does not present all PTMs or all risks. It serves to highlight salient aspects of the need for standardisation and thereby improve safety and security when using pre-trained AI models for use in healthcare. The 5-question short-quiz discussed in this paper is limited in scope and reach and was used as an early indicator only, on which further research can follow.

Acknowledgments. This research is supported by the Research Ireland under the Regulatory Compliance Framework for Trustworthy AI Medical Device Software (Reg-Fr-AIMs) project, ID 21/FFP-A/9255.

Disclosure of Interests. The 1st and 5th authors are members of IEC and CEN-CENELEC Technical Committees for Healthcare, including TC62 Artificial Intelligence. The opinion's set out here are those of the authors' alone and are not intended to represent the TC's or organizations of which they are part. The guidelines developed in this research form part of a broader project aiming at developing Trustworthy AI for Medical Devices led by DkIT and UCD, Ireland and specifically from a PhD project on AI Change Management (AICM) processes [11]. The questions raised in relation to PTMs are presented under Appendix 1 to this paper. The questions have been evaluated against the Chinese standard on PTMs for analysis of completion. Additional guidance adopted from the Chinese guidance can be seen under *Supporting Comments*.

References

[1] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever, 'Improving Language Understanding by Generative Pre-Training', 2018. [Online]. Available: https://gluebenchmark.com/leaderboard

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', May 2019. [Online]. Available: http://arxiv.org/abs/1810.04805

[3] K. Denecke, R. May, and O. Rivera-Romero, 'Transformer Models in Healthcare: A Survey and Thematic Analysis of Potentials, Shortcomings and Risks', Dec. 01, 2024, *Springer*. doi: 10.1007/s10916-024-02043-5.

[4] B. D. Menz *et al.*, 'Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross-sectional analysis', *BMJ*, vol. 384, p. e078538, Mar. 2024, doi: 10.1136/bmj-2023-078538.

[5] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition', Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.03385

[6] N. Aspell, A. Goldsteen, and R. Renwick, 'Dicing with data: the risks, benefits, tensions and tech of health data in the iToBoS project', *Front Digit Health*, vol. 6, 2024, doi: 10.3389/fdgth.2024.1272709.

[7] EU AI Act, 'EU Artificial Intelligence Act - Proposed', *Official Journal of the European Union*, no. 0106(COD), Apr. 2021, Accessed: Nov. 29, 2023. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[8] MDR 2017/745, 'Medical Device Regulation (EU) 2017/745, as amended', 2024.

[9] M. Y. Shaheen, 'AI in Healthcare: medical and socio-economic benefits and challenges', *ScienceOpen*, Sep. 2021, doi: 10.14293/S2199-1006.1.SOR-.PPRQNI1.v1.

[10] Q. Dong *et al.*, 'A Survey on In-context Learning', Oct. 2024, [Online]. Available: http://arxiv.org/abs/2301.00234

[11] N. S. J. Lynch, R. Loughran, M. McHugh, and F. McCaffrey, 'Trustworthy Artificial Intelligence in Healthcare: A Proposed Framework', in *Systems, Software and Services Process Improvement*, M. Yilmaz, P. Clarke, A. Riel, R. Messnarz, M. Zelmenis, and I. A. Buce, Eds., Cham: Springer Nature Switzerland, 2026, pp. 69–90.

[12] NMPA, 'YY/T 1833.5 人工智能医疗器械 质量要求和评价 第 5 部分：预训练模型 Artificial intelligence medical device-Quality requirements and evaluation-Part 5: Pre-trained models', 2024. Accessed: Sep. 19, 2025. [Online]. Available: https://www.codeofchina.com/standard/YYT1833.5-2024.html

[13] The Consumer Technology Association (CTA), 'Letter to FDA on Pre-Trained Model Regulation'. [Online]. Available: https://www.fda.gov/advisory-committees/advisory-

[14] ISO13485:2016, '13485', 2016. [Online]. Available: www.ili-info.com

[15] A. Vaswani *et al.*, 'Attention Is All You Need', in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA USA, 2017.

[16] X. Zhao, Q. Zhao, and T. Tanaka, 'EpilepsyLLM: Fine-tuning large language models for Japanese epilepsy knowledge representation', *Artificial Intelligence in Health*, vol. 0, no. 0, p. 025180042, Sep. 2025, doi: 10.36922/AIH025180042.

[17] M. Gräf *et al.*, 'Comparison of physician and artificial intelligence-based symptom checker diagnostic accuracy', *Rheumatol Int*, vol. 42, no. 12, pp. 2167–2176, Dec. 2022, doi: 10.1007/s00296-022-05202-4.

[18] Ada Health Corporation, 'Ada Health'. Accessed: Nov. 27, 2025. [Online]. Available: https://about.ada.com/life-sciences/

[19] S. Gilbert, M. Fenech, M. Hirsch, S. Upadhyay, A. Biasiucci, and J. Starlinger, 'Algorithm change protocols in the regulation of adaptive machine learning-based medical devices', Oct. 01, 2021, *JMIR Publications Inc.* doi: 10.2196/30545.

[20] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition', Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.03385

[21] K. Simonyan and A. Zisserman, 'Very Deep Convolutional Networks for Large-Scale Image Recognition', Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.1556

[22] M. Musthafa, T. R. Mahesh, P. V. V. Kumar, and S. Guluwadi, 'Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50', *BMC Med Imaging*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12880-024-01292-7.

[23] B. Koonce, 'ResNet 50', in *Convolutional Neural Networks with Swift for Tensorflow*, Apress, 2021, pp. 63–72. doi: 10.1007/978-1-4842-6168-2_6.

[24] K. Simonyan and A. Zisserman, 'Very Deep Convolutional Networks for Large-Scale Image Recognition', *ICLR*, Apr. 2015, [Online]. Available: http://arxiv.org/abs/1409.1556

[25] Q. Guan *et al.*, 'Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: A pilot study', *J Cancer*, vol. 10, no. 20, pp. 4876–4882, 2019, doi: 10.7150/jca.28769.

[26] P. L. Chong *et al.*, 'Integrating artificial intelligence in healthcare: applications, challenges, and future directions', *Future Sci OA*, vol. 11, no. 1, Dec. 2025, doi: 10.1080/20565623.2025.2527505.

[27] EN-62304, 'Medical device software: software life-cycle processes', Nov. 30, 2015, *CENELEC, Geneva*.

[28] E. Kopacheva, A. Henriksson, H. Dalianis, T. Hammar, and A. Lincke, 'Fine-tuning Clinical Language Models to Identify Adverse Drug Events in Clinical Text: Machine Learning Approach (Preprint)', *JMIR Form Res*, Sep. 2025, doi: 10.2196/71949.

[29] Microsoft Foundry, 'Foundation models for healthcare AI', Nov. 2025. Accessed: Nov. 25, 2025. [Online]. Available: https://learn.microsoft.com/pdf?url=https%3A%2F%2Flearn.microsoft.com%2Fen-us%2Fazure%2Fai-foundry%2Ftoc.json%3Fview%3Dfoundry-classic

[30] R. Luo *et al.*, 'BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining', Apr. 2023, doi: 10.1093/bib/bbac409.

[31] E. Alsentzer *et al.*, 'Publicly Available Clinical BERT Embeddings', Jun. 2019, [Online]. Available: http://arxiv.org/abs/1904.03323

[32] NVIDIA, 'NVIDIA H100 Tensor Core GPU.', 2024. [Online]. Available: www.nvidia.com/h100

[33] K. Lekadir *et al.*, 'FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare', *BMJ*, 2025, doi: 10.1136/bmj-2024-081554.

[34] F. Yousefi *et al.*, 'Opportunities, challenges, and requirements for Artificial Intelligence (AI) implementation in Primary Health Care (PHC): a systematic review', *BMC Primary Care*, vol. 26, no. 1, Dec. 2025, doi: 10.1186/s12875-025-02785-2.

[35] P. Slattery *et al.*, 'The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence', Aug. 2024. Accessed: Aug. 15, 2024. [Online]. Available: https://cdn.prod.website-files.com/669550d38372f33552d2516e/66bc918b580467717e194940_The%20AI%20Risk%20Repository_13_8_2024.pdf

[36] N. St John Lynch, R. Loughran, M. McHugh, and F. McCaffrey, 'Artificial Intelligence-Enabled Medical Device Standards: A Multidisciplinary Literature Review', 2024, pp. 112–130. doi: 10.1007/978-3-031-71139-8_8.

[37] EU, *The AI Act 2024/1689*. Brussels, 2024, pp. 1–144. Accessed: Apr. 08, 2024. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689

[38] CEN/CLC/JTC21, 'Information technology - AI - AI Concepts and terminology - Amendment 1: Generative AI - draft amendment out for comment only', *IEC*, Sep. 2025, Accessed: Sep. 24, 2025. [Online]. Available: https://isolutions.iso.org/ballots/part/npos/ballotAction.do?method=doView&id=61809

[39] EN/ISO/IEC-22989, 'Information Technology - Artificial Intelligence - Artificial Intelligence Concepts and Terminology', 2023.

[40] ISO/IEC-DIS-42001, 'Information Technology-Artificial Intelligence-Management system', 2022.

[41] K. Acharya, A. Velasquez, and H. H. Song, 'A Survey on Symbolic Knowledge Distillation of Large Language Models', *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 12, pp. 5928–5948, 2024, doi: 10.1109/TAI.2024.3428519.

[42] B. Zhang, L. Liu, M. H. Phan, Z. Tian, C. Shen, and Y. Liu, 'SegViT v2: Exploring Efficient and Continual Semantic Segmentation with Plain Vision Transformers', *Int J Comput Vis*, vol. 132, no. 4, pp. 1126–1147, Apr. 2024, doi: 10.1007/s11263-023-01894-8.

[43] Y. Liu *et al.*, 'RoBERTa: A Robustly Optimized BERT Pretraining Approach', Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.11692

[44] J. M. Pérez *et al.*, 'pysentimiento: A Python Toolkit for Opinion Mining and Social NLP tasks', Jul. 2024, [Online]. Available: http://arxiv.org/abs/2106.09462

[45] A. M. García *et al.*, 'Toolkit to Examine Lifelike Language v.2.0: Optimizing Speech Biomarkers of Neurodegeneration', *Dement Geriatr Cogn Disord*, vol. 54, no. 2, pp. 96–108, Apr. 2025, doi: 10.1159/000541581.

[46] N. St John Lynch, R. Loughran, M. McHugh, and F. McCaffrey, 'Artificial Intelligence-enabled Medical Device Standards: a multidisciplinary literature review', Munich: EUROSP'24, Sep. 2024, pp. 1–12.

[47] A. F. Ansari *et al.*, 'Chronos: Learning the Language of Time Series', Nov. 2024, [Online]. Available: http://arxiv.org/abs/2403.07815

[48] X. Shi *et al.*, 'Time-MoE: Billion-Scale Time Series Foundation Models with Mixture of Experts', Feb. 2025, [Online]. Available: http://arxiv.org/abs/2409.16040
[49] L. Nochumsohn, R. Marshanski, H. Zisling, and O. Azencot, 'Super-Linear: A Lightweight Pretrained Mixture of Linear Experts for Time Series Forecasting', Sep. 2025, [Online]. Available: http://arxiv.org/abs/2509.15105
[50] IEC-63521, 'Machine Learning-enabled Medical Device - Performance Evaluation Process (Draft Out for Public Comment)', May 2025. [Online]. Available: www.bsigroup.com
[51] D. Hintersdorf, L. Struppek, and K. Kersting, 'Balancing Transparency and Risk: The Security and Privacy Risks of Open-Source Machine Learning Models', Aug. 2023, [Online]. Available: http://arxiv.org/abs/2308.09490
[52] D. Hintersdorf, L. Struppek, and K. Kersting, 'Balancing Transparency and Risk: The Security and Privacy Risks of Open-Source Machine Learning Models', Aug. 2023, [Online]. Available: http://arxiv.org/abs/2308.09490
[53] FDA, 'Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission (draft) Recommendations', Rockville MD, Jan. 2025. [Online]. Available: https://www.fda.gov/vaccines-blood-biologics/guidance-compliance-regulatory-information-
[54] M. Mitchell *et al.*, 'Model cards for model reporting', in *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, Inc, Jan. 2019, pp. 220–229. doi: 10.1145/3287560.3287596.
[55] M. Elnawawy, M. Hallajiyan, G. Mitra, S. Iqbal, and K. Pattabiraman, 'Systematically Assessing the Security Risks of AI/ML-enabled Connected Healthcare Systems', Apr. 2024, [Online]. Available: http://arxiv.org/abs/2401.17136

Appendix 1: Pre-Trained Model Evaluation Checklist

Type	Consideration[1]	Supporting Comment[2]
Transparency (including Explainability as a sub-set of transparency)	Is the AI model under development utilising a 'pre-trained model (PTM)' or 'foundational' model?	PTMs refer to those used in the development of AI medical devices. They are not limited to models specifically developed for medical devices. Sources of PTMs are diverse and may include: Medical device manufacturers, Third-party suppliers, Third-party service platforms, open-source networks, the types of PTMs, which may vary in modality and parameter scale.
	Has the provider of the PTMs described and/or provided all relevant details concerning the model itself (architecture, layers, etc.), including the processes and methods used, and the datasets involved in training?	Version identification (release version number), model architecture, training data and dataset construction methods, annotation methods, training process, model applicability, etc. Note: This information can be provided as part of a Model Card, where suitable.
	Describe the existing architecture and the new architecture that will be developed; examine differences.	The documentation should describe the overall structure and origin of the model. If the structure is a modification of a publicly available one, the differences must be detailed. If it is a proprietary structure, a detailed mathematical or structural description is required.
	Provide documentation with information about model nodes and functions; learning method.	Number of nodes, neurons, distribution, activation functions, connection methods. Note: For non-neural network algorithms, describe the algorithm type and its basic units.
	Describe model parameters, hyperparameters and relationship to architecture.	The documentation should describe metadata for parameters in each layer, including number of nodes/neurons, relationships, categories.
	Describe the how the data is represented in the model and how it mirrors the model under development. Identify if changes are required to the model in development for the new use case.	The documentation should describe how data is represented in the model, including: level of abstraction, data types and scale, encoding methods (e.g., one-hot, label encoding), pre-processing and post-processing steps.
	Describe data modality; what level of documentation is provided by 3rd party supplier?	The documentation should describe the data modalities (type or form of data) used for training the PTM, such as speech, natural images, medical images, etc. If applicable, the medical data acquisition equipment should be specified. If simulation data was used for training, the data generation methods and procedures should be described in detail. If applicable, information about data annotation should be included, such as the source, quantity, and quality of annotations

[1] *Considerations are sourced from 1st Author's AICM process in development as part of larger project Trustworthy AIeMD [11]*
[2] *Developed from AICM and review against YY/T 1833.5 Chinese Standard for completeness* [12]

Type	Consideration[1]	Supporting Comment[2]
	Data volume and breakdown; what level of documentation is provided by 3rd party supplier?	The documentation should specify the total amount of data used to train the PTM. If applicable, the volume of each data modality should be described. The data partitioning method used during pre-training should be explained, including the volume of training, validation, and test sets, as well as any distribution differences among them.
	Data quality; what level of documentation is provided by 3rd party supplier?	The documentation should describe the quality of the data used for training the PTM. Examples include accuracy, completeness, and consistency. Accuracy may not be a relevant metric for healthcare activities. For example, cancer detection relies highly on sensitivity (recall) to ensure a positive case is not missed and that over-burdening the healthcare system still further does not occur through false positives. Sensitivity is better when predicting rare events like some cancer detection. Where datasets are imbalanced, metrics such as AUC-ROC or F1 score are important in understanding performance across patient groups. AUC-ROC are discriminative across thresholds but insensitive to prevalence. The criteria used to select training data should be explained, such as data source, reliability, relevance, and data cleaning procedures. It is important to align predicted probabilities with observed outcomes for critical clinical decisions.
	Describe the interpretability of the model and the level of additional development required to meet interpretability and transparency objectives.	Interpretability is essential for ethical and regulatory transparency. It helps identify biases and ensures fairness in model decisions. Consider the following: the model's internal mechanisms, relationship between spatial outputs and internal features, ability to extract meaningful information from learned features.
Security & Data Protection	Does the data comply with data management procedures/ standards?	Yes/No. Provide reference to Data Management SOP and provide description as to how the PTM complies with existing procedures/standards or actions taken to address any deficits.
	Data privacy. What level of documentation is provided by 3rd party supplier?	If applicable, the documentation should describe the technical measures used to protect subject privacy, such as data de-identification and anonymization. The rules for these privacy protection methods should also be described.
	Baseline Performance; what level of documentation is provided by 3rd party supplier?	The documentation should describe the baseline performance of the PTM on the source task, including results on the training, validation, and test sets.
	Original source tasks used to Train the PTM; what level of documentation is provided by 3rd party supplier?	The documentation should describe the source task used for training the PTM and the learning method adopted
	Training Settings, Test Settings and Configuration; what level of documentation is provided by 3rd party supplier?	The documentation should describe the data augmentation methods, model weight initialization strategies, optimizers used, and the configuration of key hyperparameters during training.
	Task Domain; what level of documentation is provided by 3rd party supplier?	The documentation should describe the task domain of the source task used during the generation of the PTM. If multiple source tasks were used, all corresponding task domains should be described.
	Model Applicability; what level of documentation is provided by 3rd party supplier? Describe original model training, fine-tuning necessary and subsequent development, testing necessary to achieve intended purpose.	The documentation should describe the data modalities applicable for forward inference using the PTM. If applicable, it should also describe the data modalities that can be processed after fine-tuning the model. Note 1: PTMs may be designed for general-purpose data modalities or specifically for medical images (e.g., MRI, CT scans) or medical text (e.g., Electronic Health Records). Note 2: Forward inference refers to the process of applying a trained machine learning model to specific tasks, such as segmentation, classification, or enhancement in medical decision support.
	Task Type; what level of documentation is provided by 3rd party supplier? How does this meet the intended purpose and ability to make generalisations in the intended population?	The documentation should describe the types of tasks suitable for forward inference using the PTM (e.g., image segmentation, image classification, signal recognition), the similarity between source and downstream task domains, the scenarios and sample formats the model can handle, and the expected performance. If applicable, it should describe robustness and generalization testing methods for downstream tasks.
	Environment; what level of documentation is provided by 3rd party supplier? Use environment (research, clinical use, etc.), technological requirements, libraries, hardware, etc.	The documentation should describe the hardware resources and software environments suitable for model inference and training. Note: This includes computing power and resources required for running the PTM (e.g., CPU, GPU, server nodes), measured by metrics such as FLOPS or operations per second. It also includes computing and software requirements for fine-tuning and deployment in AI medical devices.
	Downstream Training required; what level of documentation is provided by 3rd party supplier? of pre-training and what additional training is necessary? What level of bias can be introduced from pre-training on wider/broad sample of data?	The documentation should describe the data requirements for transferring the PTM to a new downstream task domain, especially when the output space mapping changes. The output space mapping is a set of all possible outputs a model or system can produce (e.g., set of all possible diagnoses). This should include the required dataset (volume, modality, etc.). If applicable, it should describe the training settings used during fine-tuning, such as data augmentation and training methods and any potential for bias from the original training prior to transfer.
	Assess intended purpose for the PTM; what level of documentation is provided by 3rd party supplier?	A technical report should be generated to serve as evidence for verifying the quality of the PTM.
	Privacy Protection; what level of documentation is provided by 3rd party supplier? What	The provider should declare the privacy protection measures adopted by the model, meeting the following requirements: Use appropriate techniques (e.g., differential privacy) to prevent leakage of training data, including distribution

Type	Consideration[1]	Supporting Comment[2]
	additional measures are necessary if assuming zero trust principle?	and individual data inference. Ensure protective measures are in place for data upload and storage operations generated by the model code.
	General methods for Quality Compliance of PTMs; What additional requirements/procedures are necessary to ensure PTM is adequately evaluated for its intended purpose?	The quality evaluation of PTMs includes assessment of the model description, quality characteristics, and other relevant aspects. The model provider should submit the PTM itself, its documentation, and other necessary materials for evaluation.
Quality	Trainability; what level of documentation is provided by 3rd party supplier? Consider compatibility of methods used in pre-training and fine-tuning; and underfitting/overfitting.	The provider of the PTM should declare the model's trainability indicators and provide evidence. If applicable, indicators such as loss function values and the degree of fit to the target data distribution should be used. For example, the loss function refers to the method chosen for optimization (i.e., optimization objectives). This optimization objective is what guides the model during training. It is important to ensure that the objective used by the PTM is suitable for use in the fine-tuned medical model. For example, cross-entropy loss is common in classification, however, in healthcare, a need for domain-specific function, such as Dice Loss, may be needed for medical image segmentation because it handles class imbalance better when lesions are small compared to a whole image. The loss function should be explained in relation to the data characteristics (balanced/imbalanced data sets, noisy data, size of lesion if imaging is being performed, etc.). Note: Trainability refers to the model's ability to be iteratively optimized during training.
	Architectural Scalability; what level of documentation is provided by 3rd party supplier? What assumptions are drawn from pre-training and model under development as part of SDLC. What additional architectural changes are necessary for intended purpose?	The provider should declare whether the model architecture can improve inference and training efficiency by increasing computational resources. Requirements for computing power and resources should be specified, along with written evidence. If applicable, minimum and maximum hardware configurations should be described. Architectural scalability primarily refers to a model's ability to support and adapt to varying levels of computational resources. This includes the model's compatibility with different deployment environments and its ability to scale up or down based on available hardware. The description of validated deployment environments for PTM should include the following: AI Server Specifications such as details of a single AI server with specified computational capabilities and architecture; Maximum supported model parameters and structure; Server computing power; Number of integrated AI accelerator cards; Interconnection methods between accelerator cards; Types of AI acceleration processors used; Maximum Supported AI Accelerator Cards such as the number of AI accelerator cards that can be supported by the deployment set-up and the strategy used for distributed parallel model partitioning.
	Transferability; what level of documentation is provided by 3rd party supplier? What quality metrics / measures are provided and what will be taken forward into development as part of claims. What will not be taken forward? Document rationale.	The provider should declare the expected performance of the PTM after fine-tuning for downstream tasks in AI medical devices and provide written evidence. For example, performance measures should be suitable for the task. Suitability of metrics should be assessed based on risk management activities and having a clear understanding of the clinical outcomes associated with the learning methods used in pre-training as well as fine-tuning. Understanding of scientific validity (reliability concepts), technical validity and clinical validity must all be assessed.
	Model Efficiency; what level of documentation is provided by 3rd party supplier? What additional resources are necessary for intended purpose in use environment(s)?	The provider should declare the efficiency of the PTM, including inference computation load, resource utilization, and accuracy (or suitable metrics), and provide written evidence. Suggested metrics include: Computation required for forward inference. Utilization of computing power and storage. Accuracy and inference time under different hyperparameter configurations (e.g., parameter count). Relevant metrics beyond accuracy or in place of that are appropriate to the domain and learning objective.
	Result repeatability and consistency; what level of documentation is provided by 3rd party supplier? Does the repeatability remain constant when updated for intended purpose; what additional quality metrics and objectives are necessary to achieve?	The provider should declare the repeatability of the model's output, ensuring consistent results when give the same test data. Note: Consistency means the output's meaning remains the same. For quantitative outputs, values should be identical or within the same range; for descriptive text, the semantic meaning should be consistent.
	Robustness; what level of documentation is provided by 3rd party supplier? How robust is the model to meet its intended purpose?	The provider should declare the model's robustness, ensuring it can produce correct outputs on datasets with varying diversity and domain deviations. Requirements include: Performance under noisy input data. Ability to generalize to out-of-distribution (OOD) data points.
	Generalisability; what level of documentation is provided by 3rd party supplier? What additional considerations are necessary for generalisability to new intended purpose?	The provider should declare the model's generalization ability and provide written evidence. If applicable, differences between the training dataset and real-world unfamiliar samples should be analysed based on the model's intended use and environment
	Adversarial Security; what level of documentation is provided by 3rd party supplier? What additional measures are necessary if assuming zero trust principle?	The provider is encouraged to declare the model's adversarial security. If applicable, examples should be provided showing the types of adversarial attacks handled and the model's performance under such conditions.

Evaluation Methods	Evaluation of PTM Documentation shall be maintained.	The evaluation should verify the completeness and accuracy of the model description provided in the documentation, ensuring it meets the requirements outlined above. For widely used PTMs, the format of the documentation may be adapted based on actual circumstances
	Is the description of the Pretrained Model complete?	The evaluation should verify the completeness and accuracy of the model description provided in the documentation, ensuring it meets the requirements outlined in Section 4. For widely used PTMs, the format of the documentation may be adapted based on actual circumstances
	Is the trainability of the PTM adequately documented for evaluation against the intended use?	Using the training examples and settings (including hyperparameters and applicable environments) provided by the model provider, the model should be trained. The convergence curve of the loss function should be recorded to determine whether the results meet the requirements. Ensure the loss value on the training set is plotted against the loss on the validation set over training to check for overfitting.
	Is the Architecture adequately described to assess the ability to scale for its intended use?	Adjust the hardware resources and software environment used for training and deploying the model. Record changes in inference and training efficiency to determine whether the results meet the requirements of
	Is the PTM appropriate for transfer for use in development or as a use-as-is SOUP within the SDLC? Identify additional requirements to be controlled as applicable.	Using the training examples and settings (including training methods and data augmentation techniques) provided by the model provider, test the model's performance on downstream tasks. Record performance metrics to determine compliance with transferability requirements. For models tested on public datasets, evaluate their performance on downstream tasks using those datasets to determine compliance. Note 1: Performance metrics may include, but are not limited to, those listed in this checklist. Note 2: Transfer learning is typically achieved through fine-tuning, such as low-rank adaptation for large language models or supervised fine-tuning
	Has the model's efficiency been adequately assessed?	Conduct forward inference using the PTM on relevant data. Evaluate compliance with model efficiency requirements based on: a) Required computation for inference b) Utilization of computing power and storage c) Performance and inference time under different parameter configurations
	Has the original output been demonstrated to be repeatable as published by 3rd party provider with same or different intended purpose?	Use identical test cases to perform inference and verify whether the model consistently produces the same output, in accordance with Output Repeatability requirements.
	Has the PTM demonstrated robustness in line with project objectives, measures, for existing/new intended purpose?	Evaluate the model's performance across applicable task domains, scenarios, sample formats, and expected outcomes to determine compliance with Robustness requirements.
	Has the PTM demonstrated generalisability in line with existing/new intended purpose?	Use test sets that were not included in the training data to evaluate the model's ability to generalize. Determine compliance with Generalisability requirements. Methods for Generalisability assessments should be documented and a reference provided to the selected methods chosen for generalization testing.
	Has the PTM demonstrated appropriate security methods and controls?	Create test cases to verify that the model's inference results are not misled by adversarial samples. Use black-box or white-box methods to generate adversarial perturbations and test the model's resistance to such attacks. Determine compliance with Security Requirements.
	Has the data been subject to Privacy Protection and what additional measures if any are required?	Review the privacy protection measures implemented in the PTM to determine compliance with Privacy Protection requirements.
	Include subgroup combination Testing of PTM and as part of developed model per intended use.	Group cases with similar characteristics into subgroups, then combine samples from different subgroups to form a diverse test set. During testing, the model should be able to make predictions across all subgroups without significant statistical bias between them. (Specific statistical methods should be selected based on the task.). Independent and competent assessment of in-built pre-existing bias and potential for newly introduced bias should be documented.
Generalisability	Include stress sample testing of PTM and as part of developed model per intended use.	Use atypical or hard-to-classify samples from the target database to test model performance. The model should meet the baseline performance claimed by the provider even on these challenging samples
	Include natural noise sample test cases.	Select samples with high levels of natural noise from the database and test the model's recognition accuracy. The model should produce correct results despite the noise. Note: "Natural noise" is a relative concept. Depending on the task and performance requirements, appropriate metrics should be used to measure noise levels and determine thresholds.
Robustness	Faulty Sample Test; ensure tests include non-possible test cases to ensure erroneous outputs are avoided.	Include samples of other types or mix clearly incorrect samples into otherwise valid data. The model should be able to avoid being misled and should not produce erroneous results.
	Review adequacy and appropriateness of Performance Metrics used in Pre-training. Identify additional performance metrics required for subsequent testing following fine-tuning and release in healthcare.	Provide risk assessment and rationale for choice of metrics ensuring all reasonably foreseeable risks are mitigated and statistical assumptions are documented.
	Ensure appropriate Image Classification metrics are used; provide rationale for appropriateness and assessment of continuation of use based on intended use in healthcare	For PTMs used in image classification tasks, the following metrics are commonly used: Accuracy: Percentage of correctly classified images out of the total. Precision: Proportion of true positive predictions among all positive predictions. Recall: Proportion of true positive predictions among all actual positives. F1 Score: Harmonic mean of precision and recall.

Performance	Image Segmentation; provide rationale for segmentation metrics and assessment of continuation into use into intended use.	For PTMs used in image segmentation tasks, the following metrics are commonly used: Intersection over Union (IoU): Measures overlap between predicted and ground truth masks. Dice Coefficient: Similar to IoU quantifies overlap between predicted and ground truth masks. Pixel Accuracy: Percentage of correctly classified pixels.
	Text Processing; provide appropriate metrics for evaluation of PTMs and assessment of continuation into use based on intended use in healthcare.	For PTMs used in text processing tasks, the following metrics are commonly used: e.g., Precision; Recall; F1 Score.