




A European Cancer Image Platform Linked to Biological and Health Data for Next-Generation Artificial Intelligence and Precision Medicine in Oncology

Deliverable D6.4:  
**Consensus document on  
methods to assess clinical  
effectiveness**

Reference	D6.4_ EuCanImage_UM_29092023
Lead Beneficiary	UM
Author(s)	Z. Salahuddin, Henry Woodruff, Philippe Lambin
Dissemination level	Public
Type	Report
Official Delivery Date	September 29th, 2023
Date of validation of the WP leader	September 29th, 2023
Date of validation by the Project Coordinator	September 29th, 2023
Project Coordinator Signature	

EuCanImage is funded by the European Union's H2020 Framework  
Under Grant Agreement No 952103 Version log



## 1. Version Log

Issue Date	Version	Involved	Comments
23-09-2023	v1.0	Z. Salahuddin, Philippe Lambin	Initial draft
26-09-2023	v1.1	Internal review H. Woodruff	Feedback
26-09-2023	v1.1	Xènia Puig	Feedback
27-09-2023	v1.2	Philippe Lambin, Z. Salahuddin	2 <sup>nd</sup> draft according to comments
29-09-2023	Final	Xènia Puig, Oliver Díaz, Karim Lekadir	Revised and corrected final version.

## 2. Executive Summary

Through iterative (virtual and face-to-face) meetings with experts in artificial intelligence (AI), as well as final users for the respective use cases, we have produced two documents to guide the model development and support the evaluation process to ensure accurate assessment of clinical effectiveness:

- a) A first table with the preferred evaluation metrics for different AI tasks (General Classification Metrics, General Segmentation Metrics, Detection, Explainability, Uncertainty), summarized in Table 1 based on predefined requirements.
- b) A second list containing the preferred evaluation metrics for each use case (Table 2). In this specific list, we have also added the potential biases and the risk of AI failures.

## Contents

### Contents

1. Version Log .....	2
2. Executive Summary .....	2
3. List of the preferred metrics for various AI tasks - outcome of expert consensus .....	3
4. List of the preferred metrics for various use case - outcome of expert consensus .....	6
5. Conclusions .....	11
6. References .....	11



### 3. List of the preferred metrics for various AI tasks - outcome of expert consensus

It is crucial to identify the specific medical imaging task at hand, whether it falls into the categories of classification, segmentation, or detection. The selected metrics may not consistently align with the biomedical requirements. For instance, general object detection challenges are frequently approached as segmentation tasks, leading to the adoption of metrics that fail to consider the potentially crucial aspect of precisely locating all objects within the scene [1]. AI solutions are perceived as black boxes due to the underlying complex decision-making process. The explainability of AI solutions is of paramount importance. Quantitative and qualitative evaluations are necessary to ensure robust and trustworthy explanations [2, 3]. It is also important to quantify the uncertainty associated with the AI explanations and evaluate the uncertainty in a quantitative manner. In our approach, we first define the most relevant AI tasks for the project, identifying five distinct tasks:

- a) Classification,
- b) Segmentation,
- c) Detection,
- d) Explainability,
- e) Uncertainty.

We concluded that ensuring the reliability and acceptance of AI-based systems in the clinic requires robust metrics. These metrics must meet several critical requirements to be considered acceptable and effective from the different stakeholders involved. Here, we outline key prerequisites for metrics used in AI for clinical applications.

1. **Comprehensibility to Data Scientists:** Metrics should be clearly defined, well-understood, and easily interpretable by data scientists involved in AI model development and validation. This clarity is essential for designing, training, and refining AI algorithms effectively.
2. **Accessibility to Clinicians:** Metrics should not remain confined to the realm of data scientists alone. Clinicians, who are the primary end-users of clinical AI, must also comprehend and appreciate these metrics. User-friendly presentations and explanations are essential to foster trust and collaboration between data scientists and clinicians [4].
3. **Acceptance by Domain Experts:** Metrics should not be arbitrary but rooted in clinical significance. They must reflect meaningful clinical outcomes and align with domain-specific expertise [5]. Metrics that are regularly featured in key peer-reviewed papers and endorsed by experts in the field provide confidence in their clinical relevance.



4. **Regulatory Compliance:** Metrics used for evaluating AI algorithms in the clinical context should meet regulatory standards and guidelines. Compliance with regulatory requirements ensures that AI systems meet the necessary safety and efficacy standards, inspiring confidence among users and stakeholders.
  
5. **Clinical Relevance:** Metrics should directly relate to clinical outcomes or patient care [6]. They should not merely reflect algorithmic performance but translate into meaningful improvements in diagnosis, treatment, or patient management.
  
6. **Ethical Considerations:** Metrics should also consider ethical aspects, such as fairness, transparency, and bias mitigation [7]. Ethical metrics assess the impact of AI on vulnerable populations and ensure that the technology respects patients' rights and values.

In the pursuit of deploying AI tools in clinical settings, metrics play a pivotal role in evaluating the performance and acceptability of these systems. For an AI to be deemed acceptable in the clinic, metrics must meet requirements that encompass clarity for data scientists, accessibility for clinicians, acceptance by domain experts, regulatory compliance, consistency, clinical relevance, and ethical considerations. Fulfilling these requirements not only ensures the reliability of AI in healthcare but also fosters trust among stakeholders and contributes to improved patient outcomes.

We formulated the list of the most common metrics following a review of the literature and an expert discussion (Table 1).



Table 1: List of the preferred metrics for various AI tasks - outcome of expert consensus

General Classification Metrics	General Segmentation Metrics	Detection	Explainability	Uncertainty
TPR/Sensitivity/Recall	Dice Index	Intersection Over Union (IoU)	Qualitative Assessment	Negative Log Likelihood
TNR/Specificity	Surface Dice Index	Boundary Intersection Over Union (Uncertainty Aware)	Model Parameter Randomization Test	Brier's score
PPV/Precision	Jaccard Index	False Positives Per Image	Added Value of Explanations in In silico Trial	Entropy
NPV	Hausdorff Distance	FROC Curve (AUC-FROC)	Pixel Flipping	Mutual Information
Accuracy	Hausdorff Distance 95 percentile	Average Precision at various thresholds (alpha= 0.1 to 0.75)		Variance
F1 Score	Average Symmetric Surface Distance	Sensitivity at various thresholds (alpha = 0.1 to 0.75)		Calibration Curves
Balanced Accuracy	Normalized Surface Distance			
Cohen's Kappa	Modified Hausdorff Distance			
Weighted Cohert's Kappa	Average Distance (2D)			
Mathews Correlation Coefficient				
AUC Receiver Operating Characteristic Curve				
AUC Precision Recall Curve				
		<b>Potential Biases</b>		
		Clinical Bias in the representative dataset - Presence or absence of associated malignancies		
		Clinical Bias in the representative dataset - Presence or absence of certain type of tumors		



#### 4. List of the preferred metrics for various use case - outcome of expert consensus

In the second phase, we organized several online discussions with a multidisciplinary group of experts. The starting point was the list from Table 1, a description of the case, and the previously mentioned requirements. Consequently, a list of at least three metrics per use case was developed. We noticed that having evaluation metrics with high values would not necessarily translate to usable AI. Therefore, we added two extra points, even though they were not formally named in the deliverable description:

##### I. Potential biases:

As AI imaging algorithms are eventually integrated into healthcare systems, the issue of potential bias has come to the forefront of discussions. Bias in AI algorithms can manifest in various forms and sources, including patient's data (gender, race, geographical location), technical specifications (hardware used for image acquisition, imaging protocol) or even the population's co-morbidities.

**1) Gender and Race Bias:** AI algorithms can inherit biases present in the data they are trained on. If training data predominantly includes images of one specific gender or race (e.g., white European), the algorithm may perform less accurately on patients from underrepresented groups (e.g., black African) [8]. Another example might be a model trained primarily on images of male patients may not perform as well on female patients, leading to potential misdiagnosis or delayed treatment. In order to reduce such bias, Mitigation Strategies can be put into place: Diverse and Inclusive Training Data. To address gender and race bias, it is crucial to ensure that the training dataset is diverse and representative of the entire patient population. Regularly auditing the dataset for biases and adjusting it accordingly can help mitigate these issues.

**2) Geographical Location Bias:** AI algorithms may also show variations in performance based on geographical location. Differences in healthcare infrastructure, equipment quality, and disease prevalence can affect the quality and availability of training data, potentially leading to disparities in algorithm accuracy. Mitigation Strategy: Global Data Collaboration: Collaboration between healthcare institutions worldwide can help create more geographically diverse datasets. Sharing data across borders and regions allows AI algorithms to learn from a broader range of patient cases, reducing location-based bias.

**3) Hardware and Protocol Bias:** Variations in hardware and imaging protocols can introduce bias into AI-powered algorithms [9]. Different hospitals may use different imaging devices, settings, and protocols, affecting the quality and characteristics of the input data. Mitigation Strategy: Standardization, Harmonization of AI solutions, and calibration efforts should be made to standardize imaging protocols and calibrate hardware to ensure consistency in data acquisition. AI algorithms should be designed to adapt to variations in input data to maintain accuracy across different devices and protocols.

**4) Population Co-Morbidities:** Co-morbidities within patient populations can also pose challenges for AI algorithms. An algorithm trained on a specific population may struggle to accurately diagnose patients with different co-morbidities [10]. Mitigation Strategy: Comprehensive Dataset Augmentation Including a wide range of co-morbidities in the training



data can help improve algorithm robustness. Regular updates to the dataset to reflect changing demographics and disease profiles are essential to maintain accuracy.

**5) Breast Composition Bias:** Breast composition, which includes factors like breast density, is an important biomarker for breast cancer. It can vary significantly among individuals and is often linked to age and BMI (Body Mass Index). Biases related to breast composition can affect the accuracy of breast cancer detection algorithms [11]. For instance, dense breast tissue can make it more challenging to detect tumors, potentially leading to false negatives, especially in younger, denser-breasted women [12]. Mitigation Strategy: Specialized Training Data: Developing specialized training datasets that encompass a wide range of breast compositions, age groups, and BMIs can help AI algorithms adapt to these variations. Algorithms should be designed to account for differences in breast density and adapt their analysis accordingly.

**6) Tumor Size Bias:** The size of detected tumors can vary depending on the availability of screening programs and the frequency of screenings within specific populations. Some countries may have well-established and widespread screening programs, resulting in the detection of smaller lesions. In contrast, regions with limited access to screenings (e.g., remote areas) may detect larger, more advanced tumors. Mitigation Strategy: Normalize Tumor Size Data: To mitigate tumor size bias, AI algorithms should consider the historical context of screening practices and account for variations in tumor sizes within their training data. This normalization process helps ensure that the algorithm's performance remains consistent across different screening environments.

**7) Age of Patients Bias:** The age distribution of the patient population can introduce bias into AI algorithms. Some regions or healthcare systems may have a more elderly patient demographic, while others may have a younger population. Age-related differences in disease prevalence and presentation can impact algorithm performance. Mitigation Strategy: Age-Stratified Analysis: AI algorithms should incorporate age-stratified analysis to better adapt to different patient age groups. By training on and considering age-related variations in disease patterns, algorithms can improve their diagnostic accuracy across diverse patient populations.

In conclusion, addressing bias in AI imaging algorithms for medical applications requires a comprehensive approach that accounts for a wide range of factors, including breast composition, tumor size, and patient age. Developing inclusive and representative training datasets, normalizing data to account for variations, and implementing specialized analyses for specific subpopulations are essential steps in mitigating these sources of bias. As AI continues to evolve in healthcare, ongoing vigilance and adaptation will be necessary to ensure equitable and accurate diagnostic support for all patients.

## **II. Risk of AI failure:**

For each use case, we also discussed and listed the main cause of potential AI failure. This would imply that we will need to have inclusion and exclusion criteria beforehand e.g., a certain AI algorithm to detect breast cancer would not work in patients with breast prostheses. The risk of AI failure remains a critical concern in healthcare settings, particularly when confronted with a constellation of complex factors that challenge the algorithms. We discussed



the general risk of AI failure in medical imaging and explored specific scenarios where these challenges are most pronounced.

Several overarching factors contribute to the risk of AI failure are described below (mitigations are shown further below).

- 1) Differences in Image Acquisition:** Variations in image acquisition, including differences in hardware, imaging protocols, and patient positioning, can lead to inconsistencies in the input data. AI algorithms must be robust enough to accommodate these differences to ensure reliable performance across diverse healthcare settings.
- 2) Presence of Other Types of Lymph Nodes:** Identifying lymph nodes is crucial for disease staging and diagnosis. However, the presence of infectious or inflammatory lymph nodes alongside potentially cancerous ones can confound AI algorithms, increasing the risk of misdiagnosis.
- 3) Coexisting Pathologies:** In some cases, the imaged region may contain various pathologies or abnormalities unrelated to the primary condition of interest. For instance, lung cancer imaging might reveal other pulmonary diseases, complicating the algorithm's task.
- 4) Occult Lesions:** Occult lesions, which are not readily visible or identifiable through standard imaging techniques, pose a significant challenge for AI algorithms. Detecting these hidden abnormalities demands high sensitivity and sophisticated algorithms.
- 5) Aesthetic Implants in the Breast:** Breast imaging often encounters challenges when patients have breast implants. AI algorithms need to distinguish between normal breast tissue, implant-related artifacts, and potential abnormalities accurately.
- 6) Segmentation Challenges:** Automatic segmentation is a crucial step in many AI applications to avoid human errors and variability. Poor segmentation, whether due to inter- or intra-rater disagreement or algorithmic limitations, can undermine the accuracy of subsequent diagnostic steps.

### III. Mitigating the Risk

In each of the scenarios mentioned above, mitigating the risk of AI failure demands careful consideration:

- a) Synthetic data generation:** Generative AI algorithms can be used to produce synthetic samples to increase the training dataset size of a known bias to alleviate the risk of failure [13].





- b) Multi-Modal Imaging:** Leveraging multiple imaging modalities, such as combining CT and MRI data, can enhance diagnostic accuracy, especially when dealing with complex cases involving occult lesions or coexisting abnormalities.
- c) Clinical Expertise:** Collaborative efforts involving radiologists and clinicians are critical. These experts can provide valuable insights and assist in cases that require nuanced interpretation.
- d) Data Augmentation:** Expanding training datasets to include diverse cases with variations in imaging conditions and pathological complexities can enhance algorithm robustness.
- e) Post-Segmentation Quality Control:** Implementing post-segmentation quality control steps can help identify and correct segmentation errors, reducing the risk of downstream diagnostic errors.

We concluded that AI has the potential to revolutionize medical imaging, but the risk of failure remains a formidable challenge, particularly in scenarios involving complex imaging conditions and coexisting pathologies. Addressing these risks requires a multi-faceted approach that encompasses algorithmic sophistication, clinical expertise, and data-driven strategies. By continually refining AI solutions and collaborating across interdisciplinary teams, we can work towards reducing the risk of AI failure in medical imaging and improving patient outcomes.



Table 2: List of the preferred metrics for various use case - outcome of expert consensus

Use-case	Task	Metrics	Potential biases	Risks of AI failure
1. Liver diagnosis (CT)	classification (benign/malignant)	Sensitivity, AUC, Positive Predictive Value (PPV)	Age, Geography, Lesion size	Differences in the image acquisition, coexistence of HCC with other abnormalities (like hemangioma)
2. Liver Diagnosis (MRI)	classification (benign/malignant)	Sensitivity, AUC, Positive Predictive Value (PPV)	Age, Geography, Lesion size	Differences in the image acquisition, coexistence of HCC with other abnormalities (like hemangioma)
3. Colorectal metastasis detection (CT)	Detection (localisation)	Sensitivity (for lesion sizes <10mm, 10-20mm, >20mm), Positive Predictive Value (PPV), AUC, False positive per image	Age, Geography, Lesion size	Differences in the image acquisition, coexistence of metastasis with other abnormalities (like hemangioma)
4. Mesorectal lymph node metastasis identification (MRI)	Detection (localisation)	Sensitivity, Specificity, PPV, FROC	Age, Geography	Differences in the image acquisition, presence of other types of lymph nodes
	Classification (metastasis present   not present)	Accuracy, Sensitivity, AUC		
5. Therapy response prediction based on primary imaging (for staging and restaging) (MRI)	Classification (no response   partial response   complete response)	Accuracy, Sensitivity, AUC, AP	Age, Geography, Lesion size	Differences in the image acquisition, presence of other pathology in the pelvic region
6. Molecular subtype classification in invasive ductal breast carcinoma (MG)	Classification (Luminal A   Luminal B   HER2 positive   Triple negative)	Accuracy, Sensitivity, Specificity, Precision, AUC-ROC, AUC-PrecisionRecallCurve, Average Precision, F1-score, Cohen Kappa (for imbalanced)	Breast composition, Age, Geography, Lesion size	Differences in the image acquisition protocols, Aesthetic Implants, Occult lesions
7. Treatment Response Prediction (Breast MRI)	Segmentation	Dice Similarity Coefficient, Jaccard Index, Hausdorff Distance, Modified Hausdorff Distance, Average Distance	Age, Geography, Lesion size	Poor automatic segmentation, segmentation inter- and intra-rater disagreement, Image acquisition differences
	Classification (no response   partial response   complete response)	Accuracy, Sensitivity, Specificity, Precision, AUC-ROC, AUC-PrecisionRecallCurve, Average Precision, F1-score, Cohen Kappa (for imbalanced)		
8. Breast Screening (MG)	Detection (localisation)	Sensitivity, Precision, False Positives Per Image (FPPi), AUC-FROC	Breast composition, Age, Geography, Lesion size	Differences in the image acquisition protocols, Aesthetic Implants, Occult lesions
	Segmentation	Dice Similarity Coefficient, Jaccard Index, Hausdorff Distance, Modified Hausdorff Distance, Average Distance		
	Classification (normal   benign   malignant)	Accuracy, Sensitivity, Specificity, Precision, AUC-ROC, AUC-PrecisionRecallCurve, Average Precision, F1-score, Cohen Kappa (for imbalanced)		



## 5. Conclusions

Through iterative (virtual and face-to-face) meetings with experts in AI, as well as other stakeholders for the respective use cases, we have produced two tables. We also have been inspired by an excellent consensus paper [1].

- a) In the first step we defined requirements of the ideal metrics. Then we filled the first table with the preferred metrics for different tasks (General Classification Metrics, General Segmentation Metrics, Detection, Explainability, and Uncertainty) summarized in Table 1.
- b) A second list contains the preferred metrics for each use case (Table 2). We have added at least the metrics per use case. In this specific list, we have added the potential biases and the risk of AI failures.

During our iterative discussions, we realized that the metrics on their own would be insufficient to ensure an efficient AI in a clinical context and that we should be aware of potential bias and the risk of AI failure during the training, to attempt to mitigate that risk, but also during the clinical testing. We expect that each AI will have inclusion and exclusion criteria.

This list is dynamic and could be adapted based on recent literature. We are in the process of integrating these metrics into OpenEBench.

## 6. References

- [1] Maier-Hein, L. and Menze, B., 2022. Metrics reloaded: Pitfalls and recommendations for image analysis validation. *arXiv.org*, (2206.01653).
- [2] Salahuddin, Z., Woodruff, H.C., Chatterjee, A. and Lambin, P., 2022. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in biology and medicine*, 140, p.105111.
- [3] Lekadir, K., Feragen, A., Fofanah, A.J., Frangi, A.F., Buyx, A., Emelie, A., Lara, A., Porras, A.R., Chan, A.W., Navarro, A. and Glocker, B., 2023. FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *arXiv preprint arXiv:2309.12325*.
- [4] James, C.A., Wachter, R.M. & Woolliscroft, J.O. (2022) 'Preparing Clinicians for a Clinical World Influenced by Artificial Intelligence', *JAMA*.
- [5] Alowais, S.A., Alghamdi, S.S., Alsuhebany, N., Alqahtani, T., Alshaya, A.I., Almohareb, S.N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H.A., Al Yami, M.S., Al Harbi, S. & Albekairy, A.M. (2023) 'Revolutionizing healthcare: the role of artificial intelligence in clinical practice', *BMC Medical Education*, 23(689).
- [6] Ferrara, E., 2023. Fairness And Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, And Mitigation Strategies. *arXiv preprint arXiv:2304.07683*.



- [7] Kelly, C.J., Karthikesalingam, A., Suleyman, M. et al. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 17, 195 (2019). <https://doi.org/10.1186/s12916-019-1426-2>
- [8] Nazer LH, Zatarah R, Waldrip S, Ke JXC, Moukheiber M, Khanna AK, Hicklen RS, Moukheiber L, Moukheiber D, Ma H, Mathur P. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digit Health*. 2023 Jun 22;2(6):e0000278. doi: 10.1371/journal.pdig.0000278. PMID: 37347721; PMCID: PMC10287014.
- [9] Mali, S.A., Ibrahim, A., Woodruff, H.C., Andrearczyk, V., Müller, H., Primakov, S., Salahuddin, Z., Chatterjee, A. and Lambin, P., 2021. Making radiomics more reproducible across scanner and imaging protocol variations: a review of harmonization methods. *Journal of personalized medicine*, 11(9), p.842.
- [10] Kelly, C.J., Karthikesalingam, A., Suleyman, M. et al. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 17, 195 (2019). <https://doi.org/10.1186/s12916-019-1426-2>
- [11] Yala, A., Lehman, C., Schuster, T., Portnoi, T. and Barzilay, R., 2019. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology*, 292(1), pp.60-66.
- [12] Lee, C.I., Chen, L.E. and Elmore, J.G., 2017. Risk-based breast cancer screening: implications of breast density. *Medical Clinics*, 101(4), pp.725-741.
- [13] Garrucho, L., Kushibar, K., Osuala, R., Diaz, O., Catanese, A., Del Riego, J., Bobowicz, M., Strand, F., Igual, L. and Lekadir, K., 2023. High-resolution synthesis of high-density breast mammograms: Application to improved fairness in deep learning based mass detection. *Frontiers in Oncology*, 12, p.1044496.