Provisional Translation*

Report on AI-based Software as a Medical Device (SaMD)

August 28, 2023

Subcommittee on Software as a Medical Device Utilizing AI and Machine Learning of the Science Board

* This English translation of the document submitted to PMDA by the Science Board is intended to be a reference material to provide convenience for users. In the event of inconsistency between the Japanese original and this English translation, the former shall prevail. The PMDA will not be responsible for any consequence resulting from the use of this English version.

Table of contents

1. Introduction 4
2. Analysis of related trends in and outside Japan 6
2-1. Related trends in and outside Japan 6
2-2. Analysis and discussion of trends in and outside Japan 11
3. ML bias 14
3-1. Bias within the data 14
3-2. Bias in the analytical method 15
3-3. Cognitive bias 15
4. Problems related to reuse of test data for post-market learning and current status of research for problem solving 17
4-1. Necessity of performance evaluation of post-market learning 17
4-2. Examples of performance evaluation of post-market learning 17
4-3. Examples of risky re-training methods used in performance evaluation of post-market learning and how to address them 17
4-4. Evaluation with repeated use of the same test data and risk assessment on Kaggle 24
4-5. Examples of risks in SaMD development 27
5. Bias in the development of DL/AI systems using medical images (radiological and ultrasound images) 29
5-1. Problems in the research and development method 29
5-2. Problems of biases in target medical images 29
6. Current status and challenges of learning data construction through physical model simulation 31
6-1. Use of numerical simulation with ML 31
6-2. Benefits of using numerical simulation with ML 32
6-3. Points to be considered when using numerical simulation with ML 32
6-4. Future of numerical simulation and ML 33
7. Outline of the databases developed to date and issues to be noted 36
7-1. Outline of surgery video database 36
7-2. Outline of digital pathological image database 39
7-3. Outline of ECG database 43
7-4. Outline of endoscopic image database 45
7-5. Common database challenges and issues 48
8. Discussion on data for the development of SaMD using ML/DL (training data, validation data, test data) 50
9. Summary 53

List of Abbreviation

ACP	Algorithm Change Protocol
AI	Artificial Intelligence
AMED	Japan Agency for Medical Research and Development
AUC	Area Under the Curve
CNN	Convolution Neural Network
DL	Deep Learning
DICOM	Digital Imaging and Communications in Medicine
FDA	U.S. Food and Drug Administration
GMLP	Good Machine Learning Practice
IDATEN	Improvement Design within Approval for Timely Evaluation Notice
IEC	International Electrotechnical Commission
IID	independent and identically distributed
IMDRF	International Medical Device Regulations Forum
ISO	International Organization for Standardization
ML	Machine Learning
MLMD	Machine Learning-enabled Medical Device
PEMS	Programmable Electric Medical Systems
PMDA	Pharmaceuticals and Medical Devices Agency
SaMD	Software as a Medical Device
SPS	SaMD Pre-Specifications
WSI	Whole-Slide Imaging

1. Introduction

As a report of the Science Board for Artificial Intelligence (AI)-based medical device software, "Issues and Proposals on AI-based Medical Devices and Systems 2017" [1] was published in 2017 to outline the current status of AI technology as of 2017, explain the technical aspects, and present the basic issues related to the regulatory science for AI-based medical systems[1] and the controversy about the ethics and responsibilities in using AI-based medical systems.

Many AI-based medical devices have been approved since. In 2021, "Guideline of determining whether software is classified as a medical device"[2] was issued to enhance the predictability of business related to the development of Software as a Medical Device (SaMD). The Guidelines were partially revised in March 2023 [2]. While there are high expectations for practical application of Machine Learning (ML)-based medical devices characterized by post-market clinical data learning and subsequent changes in performance, it is conjectured that the development of such devices has yet to be promoted.

Meanwhile, the regulatory authorities of the relevant countries have been discussing medical devices whose performance will change after marketing. The stakeholders are engaging in a lively discussion on it.

AI-based medical devices covered by this report include various fields of biometric analysis such as medical image analysis system, electrocardiogram (ECG), and sphygmograph as well as support systems for radiotherapy treatment planning. Devices used for the field of genomic data analysis are not included. A chat service in which AI responds in a natural interactive manner was launched while this report was still in preparation. Research and development of AI-based medical systems that answer questions asked by healthcare professionals in natural language is also underway. However, the characteristics of the technology are still unexplained, and it has been pointed out that the AI may generate wrong answers. Therefore, the AI chat technology is not included in this report since it was considered premature to discuss at the moment.

This report focuses on the following real issues that have arisen since the publication of the 2017 report and summarizes the discussion held by the Subcommittee.

- Trend analysis of activities to establish medical device regulations and safety standards in and outside Japan
- Bias issues in ML
- Problems related to reuse of test data for post-market learning and current status of research for problem solving

---

1 In the "Issues and Proposals on AI-based Medical Devices and Systems 2017", systems intended to be used for diagnosis, treatment, or prevention of diseases that include AI as a component (not limited to medical devices under the Pharmaceutical and Medical Device Act) are called "AI-based medical systems."

2 SaMD regulated by the Pharmaceutical and Medical Device Act is defined as a program (a software function) intended to be used as a medical device, which may affect the life and health of the patient (or the user) if not functioning as intended.

- Current status and issues of training data construction based on a physical model simulation
- Issues related to various clinical information databases as the data sources of training data, validation data, and test data

In the trend analysis of activities to establish medical device regulations and safety standards in and outside Japan, the concept of Good Machine Learning Practice (GMLP) and the regulatory issues about performance change are discussed.

The problems of ML bias are often discussed in terms of whether the three types of data: Training data, validation data, and test data, are generally similar to the statistical properties of the patient group for practical application. In addition to these biases, other data biases that require attention in the development of AI/ML-based medical devices, e.g., usage bias, are discussed. Examples of actual biases in Deep Learning (DL) in medical image analysis are presented by citing bibliographic references.

Considering the problems in reuse of test data in post-market learning and the current research for problem solving, training data, validation data, and test data must theoretically be collected separately to prevent contamination. If an inappropriate development method is taken, such as intentionally or unintentionally biased collection of training data for the purpose of eliminating this weakness after knowing the evaluation results, there is a possibility of overfitting even if the test data was collected independently. Such overfitting and other related issues are discussed.

Also, the current situation and issues of using numerical simulation as a training data generation method are discussed, and the related issues are summarized based on the report of the Subcommittee on Computer Simulation published by the Science Board in 2021 [3].

For the challenges related to various clinical information databases as sources for training data, validation data, and test data, the common property which the databases should have and the area-specific issues in the development of AI-based medical devices based on ML concretely using surgical image database, endoscopic database, pathological image database, and ECG database are summarized.

Protection of personal information will be a major issue in the research and development when collecting training data, validation data, and test data in the clinical practice. Such a legal issue does not fit in with the scientific discussion of the Science Board and is considered outside the scope of this report. Since it is an important issue in the actual research and development, however, the challenges in the actual database development are briefly presented. The ethical requirements that must be considered at a minimum when using the valuable data for the development of medical device software are also briefly mentioned.

2. Analysis of related trends in and outside Japan

2-1. Related trends in and outside Japan

U.S. Food and Drug Administration (FDA)

In April 2019, the FDA published a discussion paper "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)" [4]. One of the objectives is thought to be to promote the development and practical application of AI capable of changing its post-market performance, particularly ML-based SaMD, and to summarize the necessary requirements. SaMD has been approved only if its performance is "fixed." In fact, however, the discussion paper states that a new framework is required for granting approval to ML-based SaMD with a characteristic feature of performance change based on continuous learning. Therefore, the discussion paper proposes to apply the Total Product Life Cycle approach, which focuses on the quality management system and post-marketing surveillance that contribute to the maintenance of safety and efficacy, to the new framework. The general principles, including introduction of the GMLP as well as quality systems, are proposed as part of the necessary framework.

The FDA recommends in the above discussion paper to develop and implement a change control plan including SaMD Pre-Specifications (SPS) that stipulate planned changes of the performance, input, and intended use and Algorithm Change Protocol (ACP) that stipulates how to check the implementation, verification, and validation of the changes specified in the SPS to control the ML-based SaMD capable of changing its performance based on continuous learning. The recommendations subsequently became more practical. The FDA issued in April 2023 a draft guidance on recommendations for marketing submission including a predetermined change control plan for the functionality of AI/ML-enabled medical devices [5].

The FDA has expected the manufacturers to achieve transparency and real-world performance monitoring based on the feedback received and the discussion about specific actions. A five-part action plan for the future development summarizing the public comments received and the proposed actions was issued in May 2021 [6].

1. Tailored regulatory framework for AI/ML-based SaMD (including the need for guidance on the predetermined change control plan)
2. GMLP
3. Patient-centered approach incorporating transparency to users
4. Regulatory science methods related to algorithm bias and robustness
5. Real-world performance and use of real-world data

The FDA has subsequently published a list of known AI/ML-enabled SaMD, held a workshop on transparency of AI/ML-enabled SaMD, and issued the GMLP guiding principles (in collaboration with Canada and the UK) to evaluate the responses to the five-part action plan. The list of AI/ML-enabled SaMD published on the FDA website was updated on October 5, 2022, for the first time in more than 1 year [7]. The comparison with the previous list revealed that 155 new products had been approved. It includes approved products with status equivalent to "partial

change" in Japanese regulatory practice, products for which only the image data improvement or modifying the format for the use on other systems, and products for which the algorithm was not specified even in the 510(k) summary. ML-based SaMD was not distinguished from traditional SaMD with a classic algorithm by product codes or other means, making it difficult to know the exact number of approved SaMD using ML as the main function.

The FDA announced in 2017 a pilot implementation of the pre-certification program (Pre-Cert) to facilitate the review of SaMD and initiated the program in January 2019 with publication of three relevant documents. The deliverable was published in September 2022. The executive summary reported the completion of the pilot implementation of the Pre-Cert and stated the development of a new regulatory framework would be further discussed based on the various issues faced during the process [8]. The following are the issues identified.

- The methodological development is still insufficient to identify low-risk devices to be exempted from the premarket review based on an organizational appraisal such as Pre-Cert.
- It is insufficient to replace the assessment of device-specific clinical performance and the cybersecurity for all devices with moderate risk.
- Device-specific assessment is crucially important particularly for high-risk devices.

Specifically, it is necessary for the promotion of Pre-Cert review to sort out the requirements for the identification of low-risk devices, to clarify the identification method and to establish an evaluation method by sorting out individual points to consider based on the characteristics of mid- to high-risk devices. However, it is difficult to develop solutions for the above issues based on the current scientific knowledge, and further discussion will be necessary.

In addition to these measures, the FDA has established the Digital Health Center of Excellence responsible for the compilation of issues related to digital health including SaMD and the future activities under the Center for Devices and Radiological Health (CDRH). However, how much the Digital Health Center of Excellence is involved in the approval review is still unknown.

Note: The above report was prepared based on the study report of the Division of Medical Devices of the National Institute of Health Sciences. See [9] for details.

U.S./Canada/UK Joint Statement

The regulatory agencies in the U.S. (the FDA), Canada (Health Canada), and the United Kingdom (the Medicines and Healthcare products Regulatory Agency; MHRA) issued a joint statement in October 2021 that proposes 10 GMLP guiding principles to ensure safety, efficacy, and quality of AI/ML-enabled SaMD [10]. The guiding principles expect the International Medical Device Regulations Forum (IMDRF), the standards organizations, and the academic societies to advance discussions based on the GMLP. The following are the specific guiding principles.

1. Multi-disciplinary expertise is leveraged throughout the Total Product Life Cycle.
2. Good software engineering and security practices are implemented.

3. Clinical study participants and data sets are representative of the intended patient population.
4. Training data sets are independent of test sets (ensuring independence).
5. Selected reference datasets are based upon best available methods.
6. Model design is tailored to the available data and reflects the intended use of the device.
7. Focus is placed on the performance of the human-AI team.
8. Testing demonstrates device performance during clinically relevant conditions.
9. Users are provided clear, essential information (ensuring transparency).
10. Deployed models are monitored for performance and retraining risks are managed.

European Union (EU)

In April 2021, the European Commission released a proposed regulatory framework for AI-based products and services in the EU market. The official name of the proposed framework is "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts [11]." The proposed framework is also applied to medical devices governed by the European Union Medical Device Regulation (MDR) and the European Union In Vitro Diagnostic Regulation (IVDR). The objectives of the proposed rules are:

- Ensure that AI systems placed on the Union market and used are safe and respect existing law on fundamental rights and Union values.
- Ensure legal certainty to facilitate investment and innovation in AI.
- Enhance governance and effective enforcement of existing law on fundamental rights and safety requirements applicable to AI systems.
- Facilitate the development of a single market for lawful, safe, and trustworthy AI applications and prevent market fragmentation.

The proposed framework was approved by the relevant committee of the European Parliament in May 2023 after a discussion about the current status. With the aim of enforcement in 2024 or later, discussions are ongoing at the European Parliament, etc.

While the proposed framework adopts a risk-based approach and intends to regulate AI-enabled SaMD as a high-risk system, the medical device organizations such as the European Coordination Committee of the Radiological, Electromedical and Healthcare IT Industry (COCIR) are opposing the direct legislation of the proposed framework. The High-Level Expert Group on Artificial Intelligence established in Europe has also stated that broad application of general policies and regulations could be harmful in certain areas. In fact, medical device industries point out that the proposed framework contains requirements that are inconsistent with the MDR and the IVDR.

Note: Refer to the study report of the Division of Medical Devices, National Institute of Health Sciences [9] for details.

South Korea

The government initiated the AI policies in 2015. The "National Strategy for Artificial Intelligence" was published in 2019 after publication of several other strategies. A number of relevant guidelines have also been published.

The "National Strategy for Artificial Intelligence" developed by the collaboration of related ministries and agencies is South Korea's national policy and vision toward the AI era. The objectives are to secure the third competitive position in the world by 2030, create economic benefits of 455 trillion won with AI, and consequently become one of the top 10 countries in terms of quality of life.

At the 2019 IMDRF Management Committee Meeting in Russia, South Korea announced that the country was implementing a deregulation for rapid market entry of AI-based products in expectation of their significant increase.

Note: Refer to the study report of the Division of Medical Devices, National Institute of Health Sciences [9] for details.

China

At the 2019 IMDRF Management Committee Meeting in Russia, China introduced a proposed domestic guidance, "Review points for decision- making medical device software using DL technology." China emphasized that the country would consider handling DL-based SaMD with a focus on the total life cycle management of the product as the FDA did. The proposed guidance was to be published in March 2022; however, the current status is unknown.

As of May 2022, nine related guidance documents have been issued.

Note: Refer to the study report of the Division of Medical Devices, National Institute of Health Sciences for [9] details.

Japan

The activities contributing to the medical device regulations include the establishment of a review working group (WG) for preparing draft evaluation indices for AI-based diagnostic imaging support systems in the project of the Ministry of Health, Labour and Welfare (MHLW) for preparing evaluation indices for next-generation medical devices and regenerative medicine products. The deliverables of the review WG that was in operation from FY2017 to FY2018 were issued as PSEHB/MDED Notification No.1219-4 (Director, Medical Device Evaluation Division, Pharmaceutical Safety and Environmental Health Bureau, MHLW) in May 2019 after reviewing the public comments and adopted as evaluation indices [12].

The formulation and revision of certification standards is progressing to transfer SaMD with a track record of approval to the certification system according to the type and the target disease in accordance with the regulatory reform plan approved by the Cabinet on June 7, 2022. In parallel, the Pharmaceuticals and Medical Devices Agency (PMDA) has started organizing information on

review points, e.g., study conditions and evaluation points necessary for efficacy/safety evaluations, and publishing the information on the website to enhance the predictability of developers [13].

The activities to provide scientific support include the research project for pharmaceutical regulatory harmonization and evaluation of the Japan Agency for Medical Research and Development (AMED), "Study of pharmaceutical regulations on SaMD using the advanced technology such as artificial intelligence" that started in 2019. In this study, the feasibility of AI-based SaMD capable of post-market learning was evaluated under industry-academia-government collaboration. As a result, a proposal for the implementation of continuous learning and performance change by manufacturer within the existing regulatory framework, particularly the "Improvement Design within Approval for Timely Evaluation Notice (IDATEN)," was compiled and submitted to the MHLW. An experimental study to identify training data factors that affect the performance of SaMD through post-market learning was also conducted. The results were incorporated into the proposal.

As a successor project to the above study, the AMED pharmaceutical regulatory harmonization and evaluation research project "Study to contribute to the performance evaluation during the post-market learning of AI-based SaMD" was started in 2022, and an experimental study to identify the points to be considered when determining the validity of the performance evaluation process has been advanced. In the future, an industry-academia-government collaboration system will be established, and draft performance evaluation guidance will be prepared based on the results of the experimental study.

Meanwhile, the results of the Health and Labour Sciences Research Grant project were compiled and issued by the MHLW in May 2023 as a guidance for approval and development based on the characteristics of SaMD [14]. Publication of more documents reporting the outcome of regulatory measures is awaited to promote the social implementation of SaMD.

IMDRF

IMDRF Artificial Intelligence Medical Devices (AIMD) WG issued "Machine learning-enabled medical devices(MLMD): Key terms and definitions" on May 6, 2022 [15]. The discussion on future policies is ongoing within the WG. The following matters will be discussed as future initiatives, and a guidance document will be issued.

1. GMLP (2023 to 2024)
2. Predetermined Change Control Plan (PCCP) (2024 to 2025)
3. Issues raised in the above discussion (from 2025)

Data management will be discussed separately in the future.

IEC/TC 62/SNAIG

IEC/TC 62/Software Network and Artificial Intelligence advisory Group (SNAIG) is an advisory group that compiles AI issues and related matters and provides guidelines for the development of standards for future medical devices.

The SNAIG proposes the following three-tiered standards applied to AI-based medical devices.

- Basic standards (AI Base)
  The standards are expected to incorporate the requirements that serve as the basis for AI, including ISO 14971 (Application of risk management), ISO 13485 (Quality management systems), IEC 62304 (Medical device software - Software life cycle processes), and IEC 62366-1 (Application of usability engineering to medical devices) required for medical devices.
- Functional standards (AI Functional)
  For example, standards related to a specific function, e.g., image analysis and waveform analysis, are conceivable.
- Standards specific to methods according to intended use or the evaluation (AI Particular)
  For example, IEC 82304-1 (health software) and other standards specific to the intended use of SaMD and IEC 60601-1 (medical electrical equipment) and other standards specific to the intended use of Programmable Electrical Medical Systems (PEMS) are conceivable.

Although no international standard of AI Base or AI Functional is available, the SNAIG recommends an addition of the following three viewpoints to the Appendix of the standards for AI-based medical devices developed by the IEC/TC 62, considering the facts that standards including requirements similar to AI Base and AI Functional have been developed or proposed in each country,

1. Basic requirements (AI Base-related: Bias control, test methods, etc.)
2. Functional requirements (AI Functional-related: Image analysis, waveform analysis, etc.)
3. Individual requirements (AI Particular-related: X-ray image analysis in the area of dentistry, tongue image analysis, etc.)

IEC/TC 62 will initiate a Preliminary Work Item (PWI) to determine what statistical method can be used to provide the number of pages of independent training data and test data to achieve the claimed performance.

Meanwhile, the SNAIG encourages to include development of data management standards in ISO/TC 215 for health informatics and include revision of ISO 13485 reflecting the characteristics of AI-based medical devices in ISO/TC 210 for medical devices in general.

In addition, development of AI-related standards in International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC) has begun for use in the implementation of the aforementioned AI Act in Europe, presumably as a result of its implementation in 2024. Medical device-related standards may likely be proposed and developed one after another by the time this report is published.

2-2. Analysis and discussion of trends in and outside Japan

The regulatory status in the respective countries varies depending on whether AI is considered fundamentally good or fundamentally evil. The countries wishing to promote the development of

AI technologies seem to think AI is fundamentally good. The EU, on the other hand, is wary of the fundamental evilness of AI and tries to regulate the AI application to SaMD as well as to various other technologies. The EU seems to understand that excessive regulations may interfere with technological development and its benefits but, at the same time, it seems to be more concerned about the risk of human rights violations due to leakage of personal information or excessive innovation. One of the reasons for the concerns may be the uncertainty of the internal algorithm of existing AI, which may obscure the responsibility. The background of such concerns should also be common to the countries wholeheartedly believing in the fundamental good of AI. The difference is only in how the country is going to address and solve the concerns. Therefore, it is important to solve the scientific issues as well as the social issues. Since the Subcommittee was established to advance discussions to develop an appropriate science-based regulatory framework for ML-based SaMD, however, this report only calls to raise awareness of the necessity to resolve the social issues.

The programs using the classic algorithm are often not distinguished from the ML (including DL)-based programs that are the current center of attention, and "AI" is often used as a comprehensive term in the relevant countries. While there may be no problem in using "AI" as a general term, it is necessary to define AI as an ML application program considering various issues that have arisen in the medical device regulations that need to be solved.

For example, the term "AI/ML" has been replacing the term "AI" assuming the use of ML in the U.S. positioned as the leader in AI technology. However, the classification of AI seems to be indefinite, e.g., no product code is available to distinguish ML application programs from others. On the other hand, the IMDRF clearly defines and uses the term "MLMD" in its published document [15]. However, the term does not seem to be actively used worldwide. The term MLMD is not used in Japan either. ML-based AI medical devices whose performance may change based on continuous learning are often not clearly distinguished from programs with the classic algorithm. In some cases, the regulatory framework to be applied to AI-based medical devices may become a subject of discussion while the differentiation is unclear. As a result, the discussion may never be settled. Unification of terminology by the ISO and the IEC promoting the international harmonization of science and technology and introduction of the terminology in the respective countries are awaited. The current discussion seems to be carried forward without defining AI on the premise that ML is used as its foundation. Even if a consensus has been reached internally on the subject of the discussion, the finalized international standards may be used differently than what is expected. Therefore, it is important to present the internal consensus to the outside parties, or to consider the "transparency" emphasized by the U.S. and Europe. In that light, the objective of the five issues proposed by the FDA in the action plan and that of the proposed EU regulations may be used as references when developing a framework in Japan. Based on these issues, the term "AI" needs to be defined first, and then ML-based SaMD, e.g., ML-based diagnostic support programs for the detection of the characteristics of the target disease in the images, needs to be identified as the subject of discussion before proceeding. Patient risk factors such as the appropriateness of test data used for learning and performance evaluation, the

presence/absence of selection bias, and the impact of different devices used need to be considered for various assessments required for the review of ML-based SaMD. See the sections below for more details.

One of the major features of ML-based SaMD is that the performance can be improved by post-market learning. The FDA seems to have tested the Pre-Cert to make products with maximized features available. It is a very interesting attempt because the FDA seems to believe AI is a fundamentally good technology. A report published in 2023 [14] extracted the issues so far identified and stated the attempt needed to be further discussed. Future development is awaited. On the other hand, with the current scientific knowledge, it is difficult to foresee the post-market performance change and the scope of expanded indications of ML-based SaMD that may be developed. It is also difficult to classify the products according to the risk of performance change, and it is unrealistic to establish uniform evaluation criteria and requirements for them. As a result, leaving the evaluation of risk and benefit associated with the post-market performance change to the manufacturer is not practical. The FDA may have concluded that it would be realistic for the regulatory authority to review ML-based SaMD every single time its performance changes. Some Japanese manufacturers may intend to commercialize their products with the characteristic features. For now, it will be realistic procedure to review ML-based SaMD every time its performance changes as in the U.S. Promotion of active use of the IDATEN is expected to accelerate the social implementation of ML-based SaMD in Japan.

3. ML bias

In statistics, the term “bias” means “how the characteristics of the sample are different from those of the population.” It is generally difficult to define a population. The population needs to be estimated from the samples. In ML, finite samples are generally used for learning and evaluation. The research and development need to be advanced with a proper recognition of the “bias,” e.g., how much the data used for learning or evaluation are dissociated from the population with the target problem, or how much the data used for learning, evaluation, and in clinical settings are different from each other. There is another term, “bias amplification,” which means the bias is further emphasized by making inferences using an ML model trained with biased training data [16].

There are other types of biases other than the “bias” related to statistical use of data. Basically, there are three types of biases: (1) Bias within the data (whether the data used for a certain purpose accurately represent the population), (2) bias in the analytical method, and (3) cognitive bias (bias in humans). These types of biases are discussed below.

3-1. Bias within the data

When pattern information such as images is classified by ML, the data should exactly represent the population as a general rule. The ML method estimates the distribution of population characteristics to be negative or positive based on the training data. When classifying the data not used in the training process, the output of classification results depends on where the unknown data are located in the characteristic distribution estimated by ML. Therefore, the data subject to classification should be consistent with the characteristic distribution of the training data. In general, however, it is impossible to use an exact representation of the population for ML. Samples collected using several sampling methods will be used. The sampling methods may include:

1. Data collected only at a specific hospital (data sampled at a specific hospital)
2. Image data collected with a specific device (data sampled by a specific imaging device)
3. Image data from a population of a specific age group or race or with a specific pathological condition (data sampled from a specific population)
4. Partitioned data obtained during methodology development (data obtained after partitioning [a type of sampling], which do not exactly represent the population characteristics)

Problems with the sampling method 1 may include an AI-based medical device trained with data obtained at a specific hospital that may not achieve a sufficient performance at other hospitals. The unsatisfactory performance is caused by the data bias between the hospital where the training data were generated and other hospitals where the AI-based medical device is used. There may be a difference in distribution between data collected at university hospitals and data collected at local clinics.

Problems with the sampling method 2 may include an AI-based medical device trained with images taken by a specific device (e.g., chest CT scans taken by a CT device of Company A) that may not achieve a sufficient performance with images taken by a different device (e.g., chest CT scans taken by a CT device of Company B). The unsatisfactory performance is caused by the data bias between the images generated by a device that generated the training data and the images generated by a device used at the hospital where the AI-based medical device is used. Careful attention is required since the difference between imaging devices is often unrecognizable by humans.

Problems with the sampling method 3 may include an AI-based medical device constructed with image data from a population of a specific age, race, gender, or with a specific pathologic condition that may not achieve a sufficient performance with data collected from other populations. The disease characteristics often differ among the countries. For example, an AI-based medical device developed in Country A may not achieve the expected performance in expanding for Country B.

Problems in the sampling method 4 may include a data bias that may occur during data partitioning in methodological development. Careful attention is required to prevent biased partitions even during the development of an AI-based medical device using N-fold cross-validation.

Biases between data to be clinically classified (real world data) and training data (laboratory data) should be minimized to the extent possible with a thorough recognition of biases in data from a sampled population. The influence of biases may be reduced by limiting the scene, the conditions, and/or the environment of use.

3-2. Bias in the analytical method

The image recognition technique is also biased. There is a "tendency to become ○○" or a "△△ tendency" involved in a certain task. A bias in training data may be caused by a data bias. The structure of the Convolution Neural Network (CNN) has been proposed in a variety of forms. According to the structure, a bias may occur in the size of the extractable structure. There may be a difference in size of tumor extracted by the CNN structure even if similar training data are used. If the methodology is based on image processing procedures, different results will be produced depending on the characteristics of the procedures. A variety of methodology and relevant hyperparameters should be tried in the development process of AI-based medical devices to minimize the biases in the method used.

3-3. Cognitive bias

There may be artificial biases, e.g., the data or the methodology is unintentionally selected. A selection bias, e.g., from which hospitals to collect data or which data to select from a large dataset for learning, may be unconsciously generated. Data selection itself is affected by the expertise of the developer who selects data. Examples may include drawing up a development plan to use only data from university hospitals or to use only data from specific countries despite prospective

overseas market expansion. There may also be biases related to labelled data generation such as image classification using ML. An AI-based medical device may be ultimately biased because of the physician's cognitive bias leading to a bias in labelled data, which are used for ML to train a classifier. For example, Reference [17] suggests a cognitive bias of physicians in identifying the affected area based on computed tomography (CT) for coronavirus disease 2019 (COVID-19). A bias may also arise from labeling based only on image findings. Other types of bias may be generated in label creation for CT scans of COVID-19, depending on whether the label creation is based on the image findings alone or PCR test results [18]. These are unconscious biases to be carefully looked out for when developing AI-based medical devices. To this end, it may be necessary to receive comprehensive training on biases before developing AI-based medical devices.

4. Problems related to reuse of test data for post-market learning and current status of research for problem solving

4-1. Necessity of performance evaluation of post-market learning

Recent AI, particularly ML-based SaMD, can be frequently retrained through continuous learning [19]. Taking advantage of the features, performance improvement through post-market re-training with the training data from the institution is expected even if the SaMD is unable to demonstrate its performance achieved at the time of approval because of domain shift [20], or variations of data characteristics among the institutions. To make the most of the features of SaMD, the IDATEN system was developed as an approval system also applied to change plans in Japan [21]. However, the system has yet to be fully used. The reasons may include possible improvement and deterioration of performance due to post-market learning [22] and concerns about risks such as catastrophic forgetting [23]. Attention should be paid to the potential risk of repeated use of the same test data when evaluating the performance after repeated retraining.

Performance evaluation after re-training is important because performance change, especially performance deterioration, must be adequately addressed. The performance evaluation may be performed in two ways: (1) Using pre-marketing test data available at the time of approval and (2) using new test data available after marketing. Evaluation using pre-marketing test data available at the time of approval is important to ensure that the performance achieved at the time of approval is maintained without any problem such as catastrophic forgetting. Evaluation using post-market test data is necessary to check the performance of the system in operation.

4-2. Examples of performance evaluation of post-market learning

Examples of evaluation may include a comparison of pre-market and post-market performance to evaluate noninferiority or equivalence [24]. Evaluation indices and acceptable ranges (margins) need to be determined for the evaluation. The endpoints may vary depending on the purpose of the SaMD. Examples may include an area under the Receiver Operating Characteristic (ROC) curve (Area Under the Curve; AUC) and accuracy for image classification, consistency of dice score or Intersection over Union (IoU) between the ground truth area and the detected area for area detection, and regression error for regression. Appropriate values should be set for the margin from a clinical point of view. Once the evaluation indices and the margin are determined, the confidence interval should be estimated using the test data to perform a test based on the relationship with the margin in a noninferiority study, for example.

4-3. Examples of risky re-training methods used in performance evaluation of post-market learning and how to address them

Pre-market performance of SaMD is generally evaluated based on a sufficient amount of test data to draw statistical conclusions. For evaluation by post-market learning, it is desirable to be performed based on a sufficient quantity of test data. When repeating re-training, it is ideal to evaluate the performance each time based on a sufficient quantity of fresh test data that are

completely unrelated to retraining. However, preparing of ground truth data (annotation data) for evaluation may require expertise and other tests. It will not be easy because of high time and economic costs. Therefore, it is highly likely the developer is forced to go with reuse of limited test data.

There is a risk in the evaluation using the same test data. The evaluation value may be different from that of fresh test data (true evaluation value) depending on the re-training method. Examples of risky re-training methods include using results of evaluation based on the same test data for the designing of ML classifier and using results of evaluation based on the same test data for a model selection process, e.g., applying multiple classifiers to the same test data to select the one with the highest performance. In all of the examples, test data are virtually incorporated into the training process, producing a bias in the evaluation value unlike when fresh test data are used for evaluation. No correct results will be obtained in a noninferiority study using evaluation values containing biases. Examples of retraining methods with a risk of bias contamination from the true evaluation value due to overfitting to the test data and how to address the risk are presented below based on the examples and solutions proposed by Dwork et al. [25]. Note that overfitting to a specific data set will result in a large difference in evaluation values between the data set and fresh test data. Overfitting can be detected by taking advantage of this phenomenon.

- Report by Dwork et al. [25]

Dwork and colleagues were inspired by the long-known Freedman's paradox [26] and performed the following two-class experiment.

Methods: First, 3 sets of 10,000 datasets with 10,000 features were generated using a normal random number generator. Each dataset was used as an independent dataset: Training set, holdout set[3], and test set.[4] In this experiment, the correct classification accuracy should be 50% because one of the labels in the two classes is randomly assigned to the data. Next, the data were used for the training of the linear classifier. The study adopted a method of repeated training using the evaluation results of the holdout set. Specifically, the training set and the holdout set were used to select features highly correlated with the correct class label (expected to have a high classification accuracy), and then features with the same sign for correlation between the training set and the holdout set was selected to design a classifier.

Results: When the evaluation value of the holdout set was biased away from the true evaluation value after repeating the classifier designing by referring to the holdout set instead of the usual method of referring to the training set alone and the selected features became 500, the classification accuracy for the training set and the holdout set was 63% ± 0.4%. On the other hand, the classification accuracy was around 50% when evaluation was made using a completely fresh test dataset (see Figure 1. A). The classification accuracy, which should have

[3] A "holdout set" is a dataset used to optimize the complexity of the model. It is also called a validation set [30].
[4] "Holdout data" and "fresh test data" are equivalent to "test data."

been around 50%, was apparently increased and biased because of selecting the features after repeated use of the holdout set, resulting in overfitting. Dwork et al. also conducted a similar experiment under the condition that some features had a correlation with the correct class label and pointed out a bias issue (see Figure 2. A).

Solution: Algorithm based on differential privacy (Thresholdout)

To address the bias issue, Dwork et al. suggest an algorithm called Thresholdout. This algorithm was developed based on the outcomes of the study on differential privacy [27]. A holdout set is used based on the differential privacy method to prevent overfitting to the holdout set.

The specific algorithm is shown in Figure 3. Figure 1. B and Figure 2. B show the results of evaluation of a classifier retrained based on the algorithm using the training set, the holdout set, and the test set. Symbols $\varepsilon_{S_h}[\phi]$ and $\varepsilon_{S_t}[\phi]$ in row 2 (a) in the algorithm in Figure 3 were calculated using the holdout set and the training set, respectively. In this row, the holdout value is returned after adding noise if the absolute value of the difference between the holdout value ($\varepsilon_{S_h}[\phi]$) and the training value ($\varepsilon_{S_t}[\phi]$) is larger than the right side value. If not, the training value is returned as the holdout value (row 2 (b), Figure 3). Intuitively, the user access is limited to the training set if the holdout value is similar to the training value. If not, the holdout set is accessible. However, noise should be added to the holdout value. This will prevent overfitting to the holdout set if it is used repeatedly to design a classifier. By retraining the classifier using the values obtained in this algorithm, the problem of the holdout accuracy (green) nearing the training accuracy (blue) in Figure 1. A and Figure 2. A was resolved as shown in Figure B in which the holdout accuracy nears the fresh test accuracy "red." In other words, overfitting does not occur even if the same test data are repeatedly used for the evaluation of classifier re-training and the classifier is designed using the evaluation results.

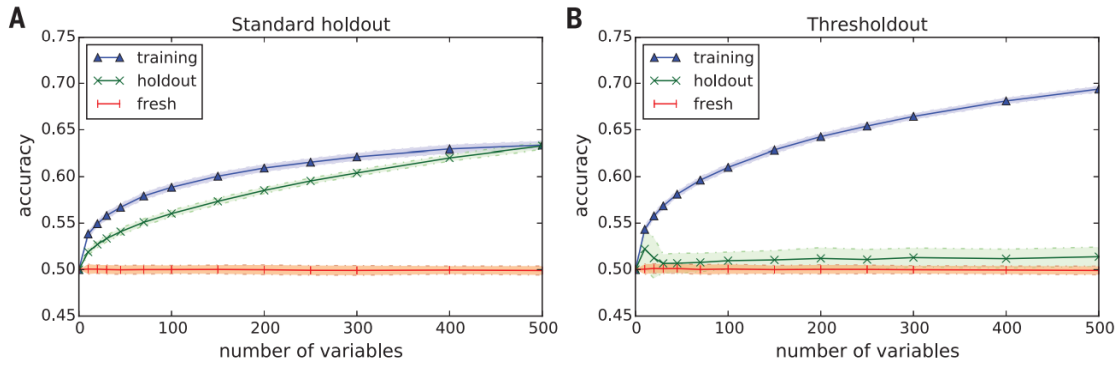


Fig.1.Learning uncorrelated labels. (reprinted from Figure 1 in Reference [25])

(A) Using the standard holdout.

(B) Using Thresholdout.

Vertical axes indicate average classification accuracy over 100 executions (margins are SD) of the classifier on training, holdout, and freshsets. Horizontal axes show the number of variables (features) selected for the classifier.

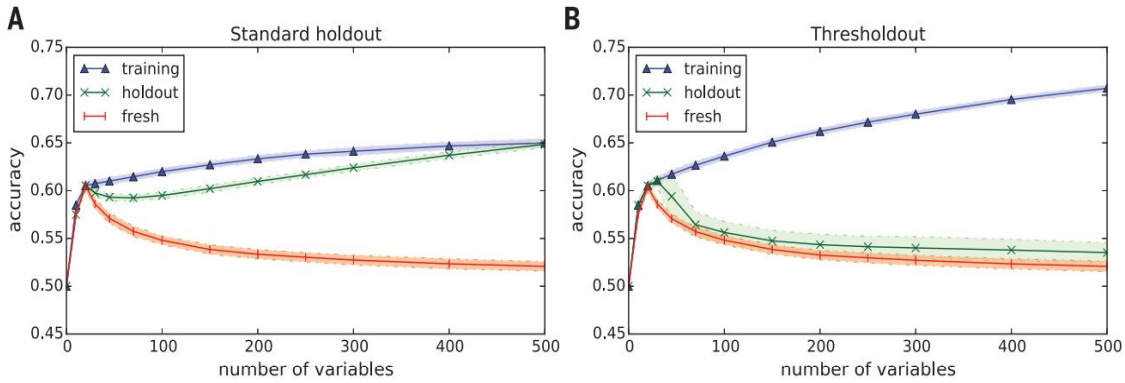


Fig.2.Learning partially correlated labels with standard holdout. (reprinted from Figure 2 in Reference [25])

(A) Using the standard holdout algorithm.

(B) Using Thresholdout.

Axes are as in Fig.1

Algorithm Thresholdout
Input: Training set S_t, holdout set S_h, threshold T, tolerance τ, budget B
Query step: Set $\hat{T} \leftarrow T + \gamma$ for $\gamma \sim \mathrm{Lap}(4 \cdot \tau)$. Given a function $\phi\colon \mathcal{X} \to [-1, 1]$, do:

1. If $B < 1$ output "$\perp$"
2. Else sample $\xi \sim \mathrm{Lap}(2 \cdot \tau)$, $\gamma \sim \mathrm{Lap}(4 \cdot \tau)$, and $\eta \sim \mathrm{Lap}(8 \cdot \tau)$
   (a) If $|\mathcal{E}_{S_h}[\phi] - \mathcal{E}_{S_t}[\phi]| > \hat{T} + \eta$, output $\mathcal{E}_{S_h}[\phi] + \xi$ and set $B \leftarrow B - 1$ and $\hat{T} \leftarrow T + \gamma$.
   (b) Otherwise, output $\mathcal{E}_{S_t}[\phi]$.

Figure 3. The details of Thresholdout algorithm.
(reprinted from Figure S1 in the supplementary material of Reference [25])

- Report by Gossmann et al. [28]

Gossmann et al. developed an algorithm using AUC as the evaluation value of SaMD called Thresholdout$_{AUC}$ [28] based on the study by Dwork et al. The algorithm is shown in Figure 4, which is almost the same as Figure 3. The main difference is that the algorithm of Dwork et al. in Figure 3 uses the correlation between the features and the labels as an evaluation value of the holdout set and the training set while the algorithm of Gossmann et al. in Figure 4 uses AUC. Gossmann et al. showed the bias was reduced in the artificial and actual data by using Thresholdout$_{AUC}$, citing the method of determining features used for classification based on the AUC evaluated with the test set[5] and the training set as an example of a risky re-training method.

Figure 5 is the result of application of the algorithm to classification of cerebral hemorrhage (presence/absence) based on the head CT scans. The figure shows the bias generated by the repeated use of the same test data to select a feature is reduced by the use of Thresholdout$_{AUC}$.

[5] The "test set" in the algorithm of Gossmann et al. and the "holdout set" in the algorithm of Dwork et al. correspond to the "repeatedly used test data."

Input: Training dataset S_{train}, test dataset S_{test}, noise rate σ, budget B, threshold T, and a stream of adaptively chosen classifiers $\boldsymbol{\phi}_i : \mathcal{X} \to [0, 1]$ for $i = 1, 2, \ldots, m$

Sample $\gamma \sim \text{Lap}(2\sigma)$ ▷ ($\text{Lap}(2\sigma)$ is the Laplace distribution with mean 0 and scale 2σ)

$\widehat{T} \leftarrow T + \gamma$

for $i = 1, 2, \ldots, m$ **do**

  if $B < 1$ **then**

    OUTPUT($\perp$)

    HALT ▷ (i.e., the test data access budget B is exhausted)

  else

    Sample $\eta \sim \text{Lap}(4\sigma)$

    if $\left|\widehat{\text{AUC}}_{S_{\text{test}}}(\boldsymbol{\phi}_i) - \widehat{\text{AUC}}_{S_{\text{train}}}(\boldsymbol{\phi}_i)\right| + \eta > \widehat{T}$ **then**

      Sample $\xi \sim \text{Lap}(\sigma), \gamma \sim \text{Lap}(2\sigma)$

      $B \leftarrow B - 1$

      $\widehat{T} \leftarrow T + \gamma$

      OUTPUT($\widehat{\text{AUC}}_{S_{\text{test}}}(\boldsymbol{\phi}_i) + \xi$)

    else

      OUTPUT($\widehat{\text{AUC}}_{S_{\text{train}}}(\boldsymbol{\phi}_i)$)

    end if

  end if

end for

Figure 4. Thresholdout$_{\text{AUC}}$ (reprinted from Algorithm 3.1 in Reference [28])

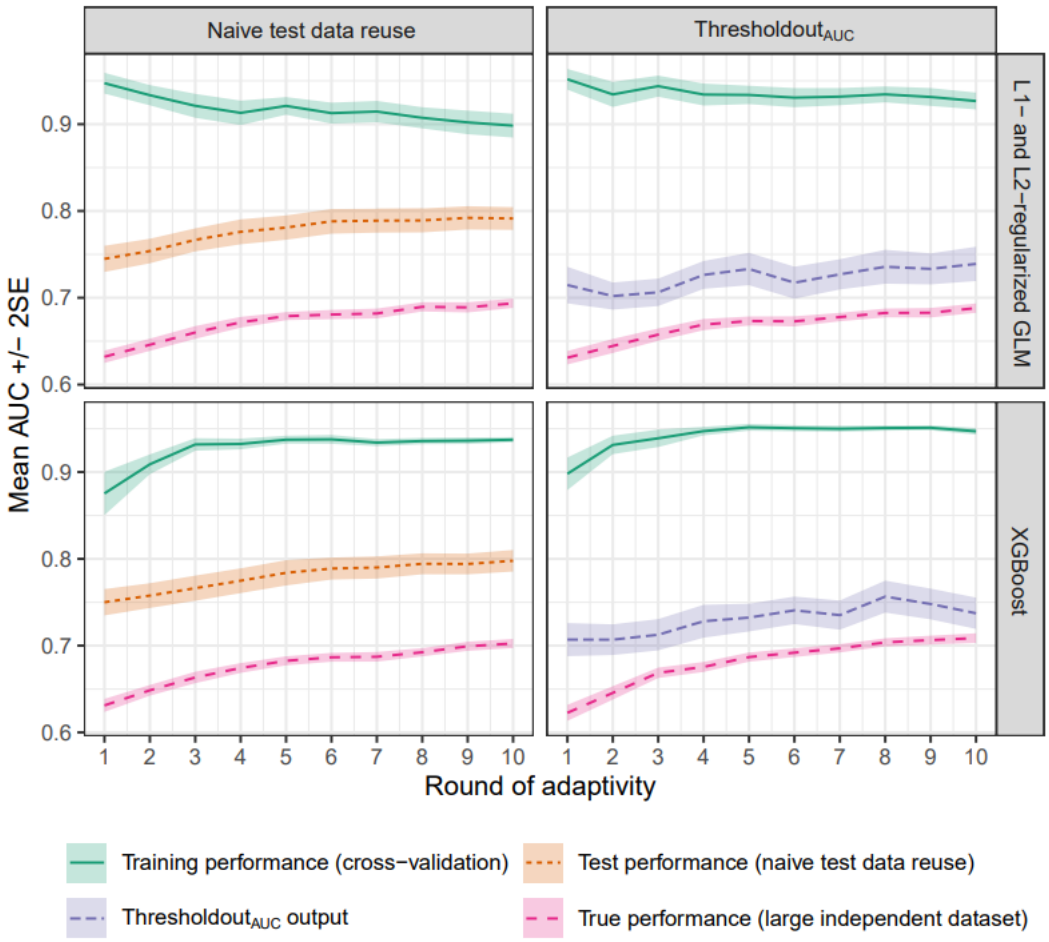


Figure 5. Change in mean AUC and standard deviation in the classification of cerebral hemorrhage (presence/absence) based on head CT scans.
The upper graphs show the results of generalized linear model; the lower graphs show the results of XGBoost. (reprinted from Figure 3 (a) in Reference [28])

Horizontal axis	: Frequency of feature selection
Green	: Performance evaluated using the training data
Orange	: Performance of repeated use of the same test data for feature selection (when Thresholdout$_{AUC}$ was not used)
Purple	: Performance of repeated use of the same test data for feature selection (when Thresholdout$_{AUC}$ was used)
Red	: Performance evaluated using fresh, large independent data

Note that Thresholdout also has some issues as an algorithm. Although bias reduction was confirmed as described above, the algorithms in Figure 3 and Figure 4 involve noise parameters τ and σ affecting the degree of bias reduction as well as a threshold that determines the accessibility to the test data $\hat{T}$. The optimal parameters and threshold will vary depending on the problem. Experimental optimization with evaluation of biases generated by overfitting to the test

data will be required to maximize the effect of bias reduction. Some issues related to accurate bias assessment, e.g., requirement of true evaluation value from fresh test data for correct bias assessment, remain to be solved to ensure proper use of the algorithm.

4-4. Evaluation with repeated use of the same test data and risk assessment on Kaggle

ML competitions in image classification and other topics have been actively held in recent years. The largest platform, Kaggle [29], offered 1,461 competitions in 2019 [30]. Kaggle competitions use scores based on two different test data. The public score will be visible on the leaderboard. The private score is used for the final ranking. Participants will be able to improve their programs again and again while watching the public score visible on the leaderboard until the competition has been closed.[6] This process is equivalent to repeating evaluation and improvement using the same test data. Evaluation using the fresh test data is considered equivalent to the private score. Although the test data are not completely fresh in a strict sense as participants receive all the test data and make a prediction, labels for the test data is not given, the order is shuffled, and it is unclear which score is public or private. Provided that these data are not revealed intentionally, evaluation based on a private score is expected to be close to evaluation based on fresh test data.

Roelofs et al. [29] selected 120 classification task competitions from among Kaggle competitions based on the conditions such as the number of participating programs (submissions) to closely compare the public and private scores. First, all the participating programs were analyzed based on the scores. The second analysis focused on the programs with top 10 percent public scores. In addition, several quantitative assessments were attempted. The public and private scores of four competitions with the largest number of participating programs (Table 1) are compared in Figure 6 and Figure 7.

Table 1. The four accuracy competitions with the largest number of submissions. (reprinted from Table 1 in Reference [29])

ID	Name	# Submissions	n_{public}	n_{private}
5275	Can we predict voting outcomes?	35,247	249,344	249,343
3788	Allstate Purchase Prediction Challenge	24,532	59,657	139,199
7634	TensorFlow Speech Recognition Challenge	24,263	3,171	155,365
7115	Cdiscount's Image Classification Challenge	5,859	53,0455	1,237,727

n_{public}: size of the public test set

$n_{private}$: size of the private test set

[6] The number of submissions accepted per day or the number of final submissions may be restricted.

In Figure 6, the public and private scores are distributed along the 45-degree straight line. No apparent overfitting to the public leaderboard data is observed. However, Figure 7 plotting the top 10 percent scores suggests overfitting to the public leaderboard data. For example, the scores are distributed below the 45-degree line in Competition 3788, suggesting a superior performance on the public leaderboard. However, the difference is not significant, about 1 pt, based on the scale on the vertical axis.

A similar analysis of 120 competitions found limited cases of suspected overfitting to the public leaderboard data; the assumption of Independent and Identically Distributed (IID) was invalid because of inappropriate division of the public and private data, the size of public and private data was small. The study concluded the public and private scores on Kaggle were comparable other than in the aforementioned cases, and the negative impact of overfitting would be limited. Another interesting point is the researchers looked into how the difference between the public scores and the private scores was related to the size of test data. The difference between the public and private scores became smaller as the size of data increased (See Figure 8).

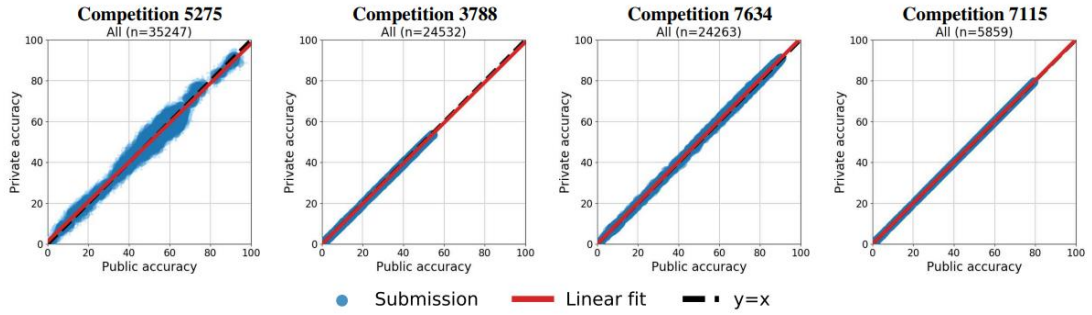


Figure 6. Private versus public accuracy for all submissions in the most popular Kaggle accuracy competitions. (reprinted from Figure 1 in Reference [29])

Each point corresponds to an individual submission (shown with 95% Clopper-Pearson confidence intervals, although the confidence intervals are smaller than the plotted data points)

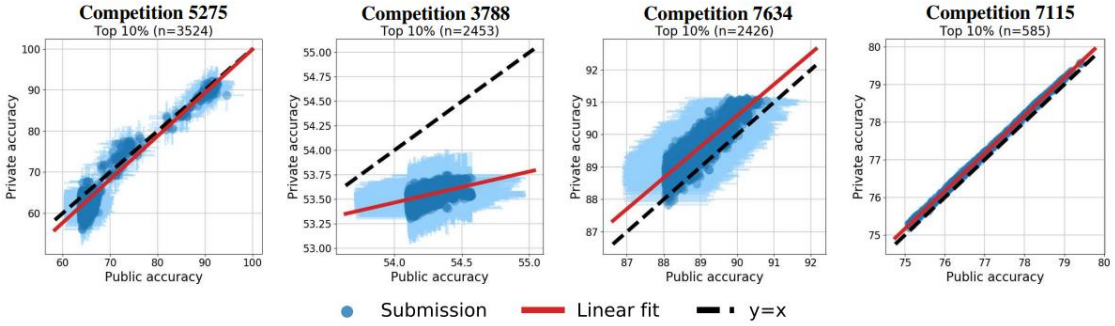


Figure 7. Private versus public accuracy for the top 10% of submissions in the most popular Kaggle accuracy competitions. (reprinted from Figure 2 in Reference [29])

Each point corresponds to an individual submission (shown with 95% Clopper-Pearson confidence intervals)

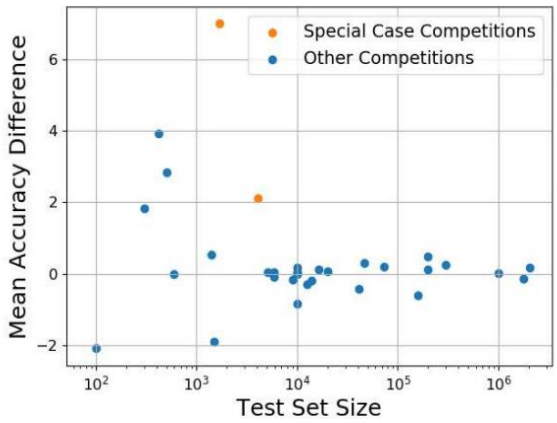
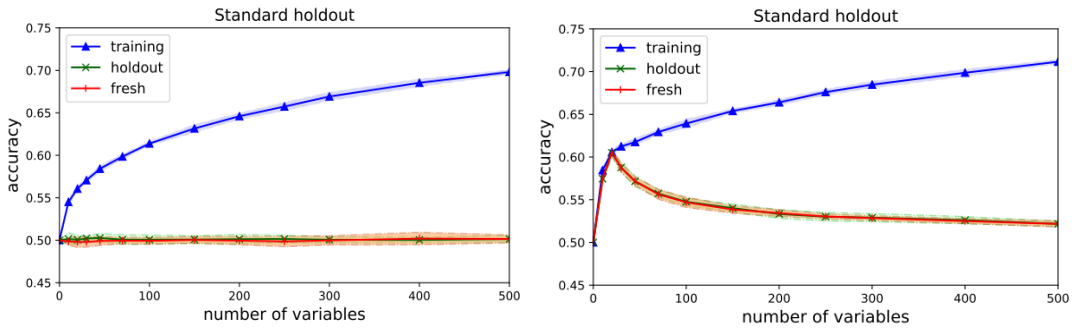


Figure 8. Mean accuracy differences versus test set size (public and private combined) for 32 accuracy competitions with at least 1,000 submissions and available test set size (the test set sizes for two competitions with at least 1,000 submissions were not available from the MetaKaggle dataset).
(reprinted from Figure 5 in Reference [29])

- Additional simulation using the data from Dwork et al.

The evaluation of the bias due to overfitting to the test data using the data from Kaggle competitions concluded the bias (difference between the public score and the private score) was small despite the repeated evaluation using the same test data in most competitions. This conclusion suggests the likelihood of bias being present may be reduced unless the "method of retraining the classifier while repeatedly using the evaluation results based on the same test data" discussed by Dwork et al. is intentionally adopted.

To confirm this against the experiments of Dwork et al. shown in Figure 1 and Figure 2 [25], an additional simulation of not using the evaluation results based on the same test data (holdout set) for retraining was conducted (see Figure 9). Although both graphs show the results of not using Thresholdout, the holdout set (green) and the fresh data (red) achieved almost the same performance. The simulation suggests a bias due to overfitting to test data will hardly occur unless an inappropriate method in which the evaluation results of test data are used for retraining is intentionally adopted. During the development of SaMD, the data at hand are randomly partitioned and used for training, validation, and test for continuous learning and repeated performance evaluation as in a cross validation. Even with such data partitioning, biases pointed out by Dwork et al. will hardly occur unless the results of evaluation based on the test data are intentionally used for training.



(A) (B)

Figure 9. Results obtained when the results of evaluation based on the holdout set were not used for retraining in the experiment of Dwork et al. shown in Figure 1 and Figure 2

(A) When the features and the correct class labels are not correlated

(B) When some of the features and the correct class labels are correlated

All results obtained when Thresholdout was not used

Vertical axis: Mean and standard deviation of classification accuracy for each dataset after 100 runs

Horizontal axis: Number of variables (features) selected by the classifier

In addition, the results of evaluation based on the same test data may possibly be unconsciously reflected in the design of the classifier. For example, the developer unconsciously uses his tacit knowledge about the design of the classifier after seeing the results of evaluation based on the same test data. Although the bias may not be significant compared with when the evaluation results are directly used as shown in 4-3 above, this may also be a risk.

4-5. Examples of risks in SaMD development

This section illustrates examples of high-risk and low-risk activities in the development of SaMD known at this point based on the aforementioned study reports and the discussions therein as well as a case of recent review by the PMDA [31]. Note that the degree of risk varies depending on the problem, that it is unknown which of the high-risk activity and the low-risk activity has a greater impact if they are performed simultaneously, and that the following examples do not cover all possible cases. Also note that only one high risk activity performed during SaMD development may result in a bias in the result of evaluation of the developed SaMD, and the bias may increase as the activity is performed more frequently.

High-risk activities in SaMD development

- Intentional use of the results of evaluation based on the same test data to the design of SaMD
  - Setting all or some of the SaMD parameters using the results of evaluation based on the test data
  - Comparing performance of multiple SaMD with different algorithms, parameters, and model structures based on the same test data to select a single SaMD

- Designing SaMD by focusing on failed test data and adding only similar data as the training data

➢ Unintentional use of results of evaluation based on the same test data for designing of SaMD
(The tacit knowledge of the developer may unconsciously be reflected in the SaMD design after an extensive analysis of the results of evaluation based on the same test data by the developer.)

Low-risk activities in SaMD development

➢ Restricting developer access to the same test data
- Physical restriction of developer access to the test data to provide a development environment where the results of evaluation based on the test data will not lead to intentional product improvement [31]
- Use of the method based on differential privacy proposed by Dwork et al. [25, 28] when reusing the results of evaluation based on the test data for an unavoidable reason
Specifically, restriction of the access to the test data, and addition of noise before use if the evaluation values based on the test data are necessary. However, it is important to appropriately set parameters of noise to be added and a threshold for determining the accessibility when this method is used.

➢ Increasing the size of test data will reduce the risk.
(For example, the meta-analysis of Kaggle competitions (Figure 8) shows a large reduction in bias in the test data with a size greater than or equal to approximately 5,000 except for some special cases.)

It is difficult to completely eliminate the risks described above. Therefore, it is important for developers and reviewers to properly understand the above risks, take steps to reduce the risks, and explain the risks scientifically and objectively.

It is also important to detect overfitting to test data caused by the risk and take actions after the detection. An example of detection method may be to prepare a test set to be repeatedly used for the evaluation of retraining and another test set for the evaluation of overfitting, and suspect overfitting when the difference in evaluation values between the datasets is statistically significant. Note that the two datasets should have an independent and identical distribution and that the quantity of data is sufficient enough to provide statistically reliable evaluation values. It is also important for the reviewer not to leak the information on the test set (including the labels) for overfitting assessment to the developer.

When overfitting occurs, the cause should be identified, and certain actions should be taken, e.g., using a low-risk development method to continue the SaMD development or preventing the problem from spreading by taking strict measures including suspension of use of the SaMD. It is important to further deepen the discussions on the issue.

5. Bias in the development of DL/AI systems using medical images (radiological and ultrasound images)

This section presents the recent debates on the bias in AI research with reference to published studies on the bias in DL/AI research using medical images.

- Problems in the research and development method
- Problems of biases in target medical images

5-1. Problems in the research and development method [32]

There are few prospective DL clinical trials or randomized clinical trials of diagnostic imaging. Most non-randomized clinical trials do not have a prospective design, involve a high bias risk, and deviate from the existing reporting standard of clinical trial outcome. Many of the existing clinical trials did not use data and codes (to be used for data preparation and modeling) and included only a small number of experts in the control group.

In general, medical imaging used in clinical trials usually involves biases such as race, gender, body constitution, and disease distribution. Needless to say, the past clinical trials were randomized (controlled) or prospective in design to reduce bias risk. This paper [32] was the result of analyzing DL research and development from the perspective of clinical trials and indicated many of the existing studies of DL-based medical imaging were not sufficiently high quality as clinical trials. Applying a prospective or randomized design required for general clinical trials to research and development of DL models as recommended by this study report may not always be appropriate considering the current stage of DL research and development.

To avoid obstacles to the approval of AI-based medical devices, appropriate clinical evaluation methods should be discussed while taking into consideration future trends in research and development of related technologies. However, efforts to scientifically reduce bias risks should inevitably be made. It will also be necessary to continue discussions on whether post-market prospective or randomized clinical trials are necessary to evaluate the performance improvement made by post-market learning.

5-2. Problems of biases in target medical images [33]

This paper [33] discusses the issue of potential bias in medical image data developed and published for other purposes as training data or test data, which is hard for developers to be aware of.

Public databases are important data sources in the current DL era; however, they may also be used “off-label” for purposes other than their original purpose. It has become common practice to use data published for a certain task to train algorithms for other tasks. The purpose of this study was to indicate the customary practice in research and development of general DL-based analysis systems may lead to overly optimistic, biased results.

According to the results, it is desirable to identify detailed imaging parameters, e.g., which vendor, image filter, and imaging conditions are used to create the images used for the training set represented by pairs of MRI signals and reconstructed images, for ML of an inverse problem solver to reconstruct MRI from MRI data. However, it may be difficult to identify all parameters for available public databases. The MRI parameters should at least be identified for test images for appropriate real-word performance evaluation. The paper suggests that it is quite possible for actual market data to be inferior to research and development test results. It is entirely possible for this phenomenon to occur not only in MRI but also in CT scans and ultrasound images. The extent to which the problem is acceptable remains controversial.

When the test results are too good in the research and development stage, whether the medical images used for research and development of AI-based medical images may have been taken under extremely unusual conditions or overly processed should also be considered. Use of medical image AI developed by 3.0T MRI for 1.5T MRI may be a simple example. Other examples may include use of AI generated from 64-row CT scans for 16-row CT scans and use of medical image AI developed with ultrasound images taken using an adaptive filter for fundamental ultrasound images.

However, it should be considered that too narrowing the adaptation by limiting the detailed parameters in the medical device review process may consequently hinder the clinical use of AI. The parameters should be appropriately set while ensuring a risk-benefit balance.

6. Current status and challenges of learning data construction through physical model simulation

Combined use of numerical simulation and ML has also been studied. ML-based medical device programs using numerical simulation in the development process or medical device programs developed based on a combination of numerical simulation and ML may become available in the near future. This section discusses points to consider when using numerical simulation in ML as of the writing of this report.

Note that the issues involved in real measurement, including biases, can be controlled with careful use of numerical simulation, but there are limitations specific to numerical simulation. The Science Board report will be a useful reference [3]. The terms used in the report are used in this section.

6-1. Use of numerical simulation with ML

The following are the primary use of numerical simulations in ML at present.

1. Training with numerical synthetic data
   "BrainWeb MRI Simulator" of McGill University is an example.
2. Training with learning data perturbed by numerical simulation
   In image ML, the machine is trained to learn images processed as in the following examples. It is a common data augmentation technique necessary for the elimination of dependence on the size, location, pixel count, etc. of the detection object. The positioning and interpretation of the use of perturbed data as test data, including its necessity, have not been established to date.
   - Deformation (rigid transformation such as rotation, parallel movement, and expansion; non-rigid transformation such as free form deformation)
   - Change in gray value (e.g., gamma conversion of concentration, random noise assignment)
   - Processing with a generative model to change the detail features while maintaining the main features of the image by assigning a noise to the feature vector of the latent space
3. Training that incorporates numerical simulation results as labels
   It has been particularly used for Generative Adversarial Networks (GAN) [34, 35]. In the medical field, Nguyen et al. combined a GAN with a finite element model to develop a "facial expression generation model based on mimetic muscle movement" to be used for prognosis prediction and rehabilitation of patients with facial palsy or facial reconstruction. The machine was trained to learn to generate smiles and symmetrical facial expressions using a model expressing tension and relaxation of individual mimetic muscles [34]. The machine compares the finite element model calculating the facial appearance (facial expression) from tension and relaxation of the mimetic muscles with the database of known facial expressions to train the generator and the discriminator. In

this example, the machine was trained without using data on the condition of mimetic muscles of patients with facial palsy.

6-2. Benefits of using numerical simulation with ML

At present, many of the numerical models available in the medical field are physical models. Whereas available physiological models are limited in type and scope. Combination with ML enables to handle problems that otherwise cannot be handled by numerical simulation alone.

In any of the methods described in 6-1 above, the numerical simulation part may often be used for statically determinate problems if they can be described as direct problems. Direct problems may involve few technical problems associated with numerical calculations, may provide stable solutions, may not require complicated processes such as solving by coupled analysis with combined equations of state, and can handled independently.

In medical practice, necessary parameters may not often be obtained from patients, or only representative values may be available. Restricting the use of numerical simulation to network training may increase the development flexibility by letting the impact of parameter ambiguity to the network side.

Use of generated data including numerical simulations may solve the ethical problem in obtaining informed consent from patients with rare diseases or fragile patients and the issues involved in handling of personal information, reduce the required real data quantity, shorten the time required for real data collection, and facilitate sharing of datasets for research and development [36].

6-3. Points to be considered when using numerical simulation with ML

The implementation process and the verification and validation (V&V) process are complicated because numerical simulation and ML interact with each other. For example, if an unintended behavior occurs, the numerical simulation and the ML must be separated to determine on which side the problem lies.

Numerical calculation using a data driven model is outside the scope of the Assessing the Credibility of Computational Modeling through Verification and Validation (ASME V&V) 40 standard. Applying the ASME V&V 40 standard to the verification of credibility of medical device simulation again to the overall process including an ML model and derivation of uncertainty quantification are unsolved issues [37].

As mentioned in the beginning, data biases may be controlled. On the other hand, the performance of numerical simulation will be limited by representable phenomena. The difficulty of handling the behavior of nonlinear systems, highly specific systems, transient response systems, and non-stationary systems increases in numerical simulation. For example, a variety of simplification (approximation) techniques will be required to deal with breakdown phenomena. The appropriateness of such approximation should be examined. Validation of the model may be necessary if lesions, pathology, or trauma are modeled in numerical simulation.

If data augmentation is performed by generating a large quantity of data while changing the parameters necessary for numerical simulation (e.g., shape, physical properties), it is necessary to evaluate the impact of parameter change and range of variation on the uncertainty of results and the calculation efficiency of numerical simulation. For example, the mesh shape and fineness determine the uncertainty of calculation results and the calculation time when using the finite element method. Locally changing the mesh size to create an extremely long and narrow mesh increases the uncertainty of calculation results. Overall size reduction to create an overly fine mesh wastes more computing time.

Although these effects can be eliminated by repeated meshing, the computing time will increase, and validation of reconstructed mesh will become necessary.

On the other hand, validation of individual data used for augmentation in all cases may not be necessary. For example, affine transformation commonly performed in medical image data augmentation prevents overfitting for a particular size, orientation, or shape by generalizing the size, orientation, and shape of the area of interest in the image. In this case, it is not necessary to seek medical validity of affine transformation of image data.

6-4. Future of numerical simulation and ML

Coexistence between numerical simulation and ML

In view of the above, when data generated by numerical simulation is used for ML, it is desirable to use it in a way that mutually compensates for the strengths and weaknesses of numerical simulation and clinical data acquisition, such as learning healthy conditions that are lacking with data obtained from patients. The model that produces facial expressions [34] simulates facial expressions with a numerical model using tension and relaxation of mimetic muscles as inputs, assigning the numerical simulator a task it is good at.

Generation of test data by numerical simulation

Data generated by numerical simulation is generally applicable to ML as data for training (including training data and validation data). On the other hand, there are challenges to overcome in using data generated by numerical simulation as test data (data used to evaluate the performance of the final ML model).

- Training data generated by numerical simulation

  The reliability and credibility of numerical simulation depend on the scheme of model used in the underlying numerical simulation and the "evidence" supporting the parameters used in the simulation (Chapter 6 [3]).

  Experimental models: numerical simulations based on data obtained from measurements on healthy subjects and animals are considered less credible compared to deductively derived models: numerical simulations based on data obtained from the target patient population.

- Test data generated by numerical simulation

  The dominant view at the moment is that the actual data should be used as test data. On the other hand, ML may be used in end-to-end clinical trials (or performance evaluation studies described in "Handling of Performance Evaluation Tests of Diagnostic Medical Devices Using Existing Medical Image Data without Involvement of Additional Invasiveness or Intervention" [PSEHB/MDED Notification No. 0929-1 from the Director of Medical Device Evaluation Division, Pharmaceutical Safety and Environmental Health Bureau, Ministry of Health, Labour and Welfare; September 29, 2021]) as well as for evaluation of specific performance indices (e.g., "Evaluation of resistance to body motion associated with image acquisition"). In such a case, the results of numerical simulation that are designed, implemented and validated with a focus on the specific performance indices may be used for the evaluation of ML performance. Specifically, Context of Use and Validation in the ASTM V&V 40 standard should be discussed in relation to performance indices to be evaluated in ML.
- Positioning of artificially processed or generated test data

  Numeric simulation results are not accepted in medical practice as being equivalent to the observed facts because of abstraction and ambiguity involved in the simulation.
  Empirically, the overall evaluation process becomes complex and difficult as more factors for ambiguous position and interpretation are involved in the development and evaluation of medical devices. The positioning and interpretation of ML application to medical devices have not been established since it is still in the developing phase. Use of numerical simulation results for the evaluation of an ML model as a medical device is likely to make the evaluation more difficult. On the other hand, numerical simulation may be used to complement the fact appropriately observed in the simulation in the field of engineering. In the evaluation of medical devices, it can be said that it is a transitional period that accepts numerical simulations.

Example of scenario for using data in numerical simulation: Learning and evaluation of resistance to body movements

Note: The following scenario was created for the explanatory purpose of this report. Its feasibility has not been tested. A recommendation of the scenario is not intended.

In the case of the development of Computer-Aided Diagnosis (CAD), which claims to be unaffected by image distortions due to body movements (motion artifact), it is necessary to collect relevant images to verify correct outputs are obtained even in the presence of motion artifacts. However, it is not easy to collect such images because patients are instructed not to move in clinical practice. Although it is possible to instruct patients to move intentionally during imaging for evaluation purposes, collecting a large number of such images may not be appropriate in terms of cost and, in some cases, ethics (e.g., it is difficult to justify radiation exposure for the purpose).

Therefore, images of numerically simulated movements are generated and used for training and test. Numerical simulation is performed and verification and validation (V&V) are conducted by the following method.

- Context of use: Data to prove that it is not susceptible to body movements are used in training and test of ML model of CAD for the detection of lesions in CT scans of lung.
- Algorithm: Based on the input static images of patients, absorbed dose measured by an x-ray detector is calculated, and scan images are reconstructed using the CT algorithm. Respiratory movements are added when resampling from the voxel data for the calculation of absorbed dose.
- Verification: The images obtained by the CT algorithm were compared with those obtained by moving the phantom in the experimental CT device to verify the deviation of the detector measurements, the shapes of depicted images and the CT values were within the predetermined acceptable range.
- Validation: A comparison was made with the images of the same moving phantom in a clinical CT device. Artifacts clinically generated by body movements were evaluated by a radiologist. As a result, the artifacts generated by the simulation under certain conditions were found to be more excessive compared with those generated by the phantom moved in the clinical CT device and the radiologist made a similar assessment.

  It was possibly caused by the correction made by the clinical CT device assuming body movements. Since the correction method is not published, it is difficult to reverse engineer it and add to the simulation.

Conclusion: When using images generated by simulation for training and test, it should be ensured specially that the model trained to ignore excessive artifacts does not disregard other important disease characteristics.

7. Outline of the databases developed to date and issues to be noted

Various novel ideas have been executed to construct databases of various medical images in Japan. To outline the individual databases, four representative databases (surgery videos, digital pathological images, ECG, and gastrointestinal endoscopy) with prospective collection of medical images under the AMED support and the relevant issues are summarized from the following perspectives:

1. Purpose of database construction
2. Type and specifications of collected data
3. Current scale of data collection
4. Type of linked medical information
5. Need for data collection in different individuals, regions, medical institutions, countries, and time periods
6. Dependence on human procedures
7. Diversity of imaging/recording devices
8. Diversity of devices used
9. Main annotation method and method of labelled data generation
10. Expected use
11. Database issues

7-1. Outline of surgery video database

Purpose of database construction

The purpose of database construction depends on the desired medical images. Gastrointestinal endoscopy (excluding treatment), pathological images, and ECG are recorded for direct diagnostic purposes while surgery videos are recorded data of the entire treatment process. In this regard, the purpose of database construction in gastrointestinal endoscopy, pathological images, and ECG is to provide annotation data aiming to assist and automate diagnosis. On the other hand, the purpose in surgery videos is often used as annotation data for objective evaluation of surgical techniques, support of procedures, and presentation of anatomical structures. Although use for educational purposes will also be considered, the quantity of data used may be limited compared with annotation data.

Type and specifications of collected data

The data to be collected include videos of laparoscopic or robotic-assisted surgical procedures. In Japan, cholecystectomy and sigmoid colectomy, for which the surgical procedure has been standardized earlier, were the first to be collected in the database.

There was no uniform standard for the collected surgery videos. The extension, scanning format, and image quality differ depending on the endoscope manufacturer and the recording device used.

Current scale of data collection

The database constructed primarily by the National Cancer Center Hospital East with support of the AMED is the largest among the existing databases in Japan. Thirteen surgical procedures and about 4000 patients are included (https://www.s-access.ncc.go.jp/). The data were collected cross-sectionally in the disease areas including large intestine (5 procedures, 36 institutions), stomach (3 procedures, 21 institutions), liver/gallbladder/pancreas (3 procedures, 26 institutions), and prostate (1 procedure, 17 institutions) from 84 institutions in total (excluding academic societies). The data were collected from various types of medical institutions including university hospitals as well as community hospitals and cancer centers, and covered the surgical procedures actually performed in the disease areas, e.g., laparoscopy, robotic surgery, and TaTME,. Data of patients with complications were also collected as much as possible to evaluate the correlation with postoperative clinical outcome. For example, about 10% of patients who underwent low anterior resection of rectal cancer and about 9% of those who underwent distal pancreatectomy for pancreatic tumor had intraoperative or postoperative complications in the database, although it depends on the surgical procedure.

Information on the technical certification by the Japanese Society for Endoscopic Surgery and the number of years of experience in endoscopy was also collected to evaluate the surgeon's technical achievement. Collection of videos of endoscopic surgery performed by beginners was also requested to evaluate what kind of differences arise in the video due to differences in the surgeons' techniques.

Type of linked medical information

The information linked to the surgery video included characteristics of the patient (e.g., gender, age, BMI, treatment history, clinical stage of malignant tumor), surgeon information (years of experience as a surgeon, endoscopic surgical skill qualification, proctor qualification for robotic surgery), device information (type of scope, type of video system, data output method), and clinical outcomes (pathological results, intraoperative complications, postoperative complications, recurrence).

Need for data collection in different individuals, regions, medical institutions, countries, and time periods

Since cancer invasion and spread vary among patients, the extent of curative resection may also vary even if the same procedure is performed. Differences in surgical proficiency level among surgeon and institutions have been pointed out. Such a bias should also be considered. Different devices (forceps, energy device, and robotic device) are generally used according to the preference of the surgeon and the institution. The trends in surgical procedures performed and devices used may change over time. Therefore, a variety of surgery videos may need to be collected depending on the objective of research and development.

Dependence on human procedures

The contents of the surgery video depend on the decision of the surgeon. Even if the standardized surgical procedure is performed, the patient's condition as well as the operation of an endoscopic camera and the method of/procedures for dissection and visual field expansion will be different.

Diversity of imaging/recording devices

There is a diversity in endoscope manufacturers (including robotic surgical devices), recording devices, and imaging modes (e.g., Narrow Band Imaging（NBI）). In particular, differences between endoscope manufacturers result in change in color tone and screen display. Also, video compression methods are different depending on the recording device.

Diversity of devices used

While many types of devices, e.g., forceps, suction tube, anastomosis device, and hemostatic agent, are used in endoscopic surgery, the number of companies related to surgical robotics are limited. Robotic surgery requires dedicated devices. As the devices for endoscopic surgery are used by not only surgeons but also assistants, 4 to 5 surgical instruments may appear at a time on a surgery video.

Main annotation method and method of labelled data generation

The surgical process annotation is created by classification. Information on the organ/surgical instruments and energization of electric scalpel is created by classification, detection, and semantic segmentation. All tasks need to be created and supervised by a surgeon. The challenge is that the border of processes or organs becomes blurred even if well defined.

Expected use

The database may be used, as a medical device, to prevent complications by assisting the surgeon in intraoperative diagnosis and navigation or for technical evaluation.

Surgery video database issues

First, there are no standardized video specifications to be used by endoscope manufacturers and recording devices. Unstandardized specifications for meta-information (e.g., scanning format, extension) of videos at the time of output makes standardization cumbersome.

Second, protection of personal information should be considered since surgery videos may show outside of the patient's body or include sounds and voices. In addition, the ID and patient name may be displayed in the operative video if cholangiography is performed during surgery and the examination video is saved in the operative video. It is necessary to carefully review the video to ensure that it does not include the personal information.

7-2. Outline of digital pathological image database

Purpose of database construction

There are two major objectives of databases related to pathological diagnosis. One is to serve as a library for educational purposes or diagnostic support. Examples include databases of typical tissue images of common diseases and databases developed for the purpose of supporting diagnosis of rare histological types such as rare cancer. The other objective is to be used as a pathological image database for the development of SaMD to support pathological diagnosis, which contains a vast quantity of general pathological images annotated for the purpose of program development.

Type and specifications of collected data

Several databases of digital histopathological images have already been developed in Japan. Digital pathological images used to be static images taken with a digital camera or a 3D camera attached to an optical microscope in Joint Photographic Experts Group (JPEG) or Tag Image File Format (TIFF) format. "Virtual slide scanner," a special digital pathological imaging device that digitizes a whole glass slide and converts it into a digital image that can be enlarged or reduced like an optical microscope, has been developed in recent years. Most existing image databases have been replaced by the virtual slide scanner images.

Digital pathological images generated by a virtual slide scanner are called Whole Slide Imaging (WSI). Many virtual slide scanners are manufactured and sold by a number of companies. However, there is no standardized WSI format; each company uses a different WSI format. The Digital Imaging and Communication in Medicine (DICOM) format is used only in a few devices sold in Japan. The volume of a WSI is large, about 200 MB to over 1 GB, as it enables automatic image creation taking into account the z-axis direction (specimen depth) to accommodate enlargement and reduction on the digital image. As with histopathological images, cytological WSIs are also available. Since thickness of cytological specimens is 4 to 5 times than an average of histological specimens (about 5 μm), more multilayering in the z-axis direction will be necessary to create cytological images compared with histological specimens. A WSI may often be over 5 GB.

Current scale of data collection

For the purpose of research and development of an AI-based pathological diagnosis support program, histopathological WSIs were collected mainly from 16 university hospitals and 7 community hospitals in the "Japan Pathology Artificial Intelligence (AI) Diagnostics Project (JP-AID)" [38] initiated by the Japanese Society of Pathology (hereinafter referred to as the JSP) in cooperation with the National Institute of Informatics, the University of Tokyo, Nagoya University, and Kyushu University with research support from the AMED. About 200,000 WSIs were eventually collected [39], and 122,000 were archived. The WSIs linked to relevant pathological diagnoses are now published exclusively to the members of the JSP on a trial basis. Arrangements are being made to make the WSIs also available to non-member researchers in the

future. In addition, as a database for educational purposes and pathological diagnosis assistant, the “Rare Cancer Database” operated by the JSP under the “Project for the Development of Pathologists for Rare Cancer Diagnosis (national grant project)” containing WSIs of rare cancers (brain tumor, bone and soft tissue tumor, pediatric tumors, lymphoma, skin tumor, head and neck tumor, and rare histological subtypes of 5 major cancers) in about 1700 patients, brief commentaries on related diseases, and 5-choice questions for E-learning has been used for pathologist certification renewal training. The project is ongoing as of 2023. About 250 to 300 rare cancer WSIs will be added to the database every year. The database also has a reverse lookup function. WSIs and commentaries related to a certain disease can be searched and browsed by entering the disease name. The database is currently open only to the members of the JSP. In the future, it will be open to non-members in Japan as well as in other countries.

Type of linked medical information

Medical information linked to WSIs includes organ name, gender, approximate age, and pathological diagnosis. The organ name and the pathological diagnosis are linked to almost all WSIs. The age of the patient may be linked to the data since age may be essential factor for pathological diagnosis of rare cancer (particularly for pediatric tumors, the diagnosis/prognosis in a 1-year-old patient may differ from a 2-year-old patient). Therefore, special attention should be paid to the handling of personal information. Information on “when the image was taken” is not linked to the WSI at the moment. However, the assignment of this information is very important. For example, the World Health Organization (WHO) Classification of Tumours is revised periodically. The name of pathological diagnosis may be changed even if the histology is the same, or the classification of a formally malignant disease may become benign by new knowledge. Therefore, the information on “when the image was taken” needs to be assigned and researchers should consider it when using WSIs. Since the WHO Classification of Tumors of the Central Nervous System has significantly changed, the aforementioned Rare Cancer Database of the JSP needs to be revised.

Need for data collection in different individuals, regions, medical institutions, countries, and time periods

Currently in Japan, pathology specimen slides are stored as “data” at most medical institutions. However, loss of the original slides means loss of the data since glass slides may break and cannot be copied. In fact, many glass slides were broken during The 2021 off the Pacific coast of Tohoku Earthquake and a number of medical institutions lost their pathological data. On the other hand, digital images can be reproduced and stored in different places. From the perspective of data storage, it is important to digitize pathological images and collect and store the digital data at the medical institution and in an external cloud server.

In global data collection, attention should be paid to “subtle differences” in the diagnostic criteria established in each country, particularly in pathological diagnosis. For example, “colonic polyp” may transition from so-called benign “colorectal adenoma” to “colorectal cancer.” How

to distinguish adenoma from cancer may be different depending on the country since it is a transitional lesion. The Japanese diagnostic criteria for cervical lesions are different from the US criteria due to the social context, including possibility of medical lawsuits. In other words, it is important to consider which country the data come from when constructing a global database. Information on the data origin needs to be assigned to collected data.

Dependence on human procedures

Microscope specimen slides stained with Hematoxylin Eosin (HE) (HE-stained slides) are used for pathological diagnosis. The HE-stained slides preparation process still often includes manual work by laboratory technicians. The tone and color gamut may be different among the institutions depending on the solution used and the staining time. Sometimes the difference may be visible. Usually, a thin tissue section of about 4 to 6 μm is sliced out with a special device called microtome to prepare a specimen slide. This slicing procedure is mostly performed manually by laboratory technicians in any country. For this reason, the thickness of the tissue section may not be within the range of 4 to 6 μm. and may often be different among the laboratory technicians performing the slicing procedure or from day to day even if the same laboratory technician performs the procedure. The HE staining process has not been standardized, making the inter-institutional difference more significant in synergy with the aforementioned difference in tone and color gamut.

Diversity of imaging/recording devices

As for imaging devices, virtual slide scanners for WSI creation are sold by about nine companies in Japan. In the United States, on the contrary, there are more than 80 companies, including small and medium-sized venture companies, selling virtual slide scanners. The format specifications of WSI vary as described earlier. Many companies have developed their own slide viewer software that meets their own format specifications. In Japan, there was an argument about the necessity of certain technical standards. The “Technical Standards for Digital Pathology System for Pathologic Diagnosis” for hardware such as virtual slide scanners and monitors has been established and published by the JSP, the Japanese Society of Digital Pathology, and the device manufacturers/distributors [40].

In addition, the regulatory approval standards for medical devices established by the MHLW are applied to virtual slide scanners. Specifically, Medical Device Nomenclature are given, “storage and display device for whole slide pathological images” as a Class I medical device and “diagnostic whole slide imaging support device” as Class II medical device. The former refers to a digital pathological imaging system that stores and displays WSIs but cannot be used for pathological diagnosis. The latter refers to a digital pathological imaging system from which WSIs are displayed on the monitor for pathological diagnosis. In clinical practice, usage of these two types of devices needs to be strictly distinguished from one another.

Main annotation method and method of labelled data generation

Annotation to WSI and labelled data generation are purely human processes [41, 42]. The processes involve two issues.

One is the difficulty and uncertainty of annotation due to "transition of the lesion" as mentioned earlier. In pathological diagnosis, the lesion may transition from benign to malignant. Although the lesion is usually determined benign or malignant based on the HE-stained slide in many patients, it may often be indeterminable. For example, some types of tumors are pathologically categorized as "benign," "malignant," or "intermediately malignant." How a tumor is categorized as "intermediately malignant" may significantly differ among pathologists. "Interobserver biases," e.g., what is determined benign by Pathologist A is categorized as intermediately malignant by Pathologist B, and then it is categorized as malignant by Pathologist C, may often be found at the time of labelled data generation. Therefore, "criteria" is required before annotation.

The second issue is the time consuming properties of labelled data generation. The lesion in a WSI is finely separated into malignant and benign regions by marking, lining, and circling as a Region of Interest (ROI) with a human hand. The process is carefully conducted since a benign/malignant differentiation is often difficult after making small patches for ML. The process takes about 15 to 20 minutes per WSI on average. It will take 66,667 hours to annotate 200,000 images, assuming annotation of a WSI takes 20 minutes. If a pathologist focuses on the annotation work for 8 hours a day, it will take about 2 and a half months to annotate all WSIs by 100 pathologists. Some paper-based reports on the research and development of AI-based pathological diagnosis support programs using unlabelled data are available. However, it has yet to reach the practical stage for pathological diagnosis support programs to be used in daily clinical practice. Technological innovation to increase the efficiency of annotation, which is a human-operated rate controlling step, will also be a major challenge in using WSI data.

Expected use

There are several types of SaMD for pathological diagnosis support awaited by pathologists. The JSP is currently requesting a listing of SaMD that "double checks" the pathological diagnosis in the next revision of medical service fees. There is a chronic shortage of pathologists in Japan. Many medical institutions have no pathologist. About 45% of the medical institutions with full-time pathologist(s) are "one-pathologist institutions" where only one pathologist is available. At one-pathologist institutions, pathological diagnoses, which are the final diagnoses of diseases, are reported without being double-checked. Establishment of an accuracy control system and reduction of psychological burden on the single pathologists have been a long-standing challenge for the JSP. For example, SaMD that double-checks pathological diagnoses of gastrointestinal endoscopic specimens, the majority of pathological diagnoses, is expected to support single pathologists by ensuring accuracy control and reducing their psychological burden. Active use of the annotated WSI database for the development of the SaMD is encouraged [41]. Although it is still in the research and development phase, some organizations are creating a gene mutation database to develop SaMD for predicting gene mutation based on HE-stained slides or a database

to develop SaMD for assuming a diagnosis based on similar images of HE-stained slides in the rare cancer WSI database and proposing immunostaining and gene mutation search for definitive diagnosis [43].

Digital pathology video database issues

The "Committee for Handling of Personal Information and Anonymized Personal Information" checked for containing of electronic metadata, etc. in the WSI collected under the support of the AMED. It took a considerable time. As mentioned earlier, WSIs come in multiple formats. Only 122,000 pathological tissue images of the stored WSIs are open to the public because a different viewer software is needed for each of the formats. Development of a standardized format and general-purpose viewer software is awaited.

7-3. Outline of ECG database

Purpose of database construction

The main purposes of ECG database construction are to improve the accuracy of automated diagnosis as a diagnostic aid and to predict a paroxysmal disease from ECG in a non-paroxysmal state. For the latter purpose, annotation data should be prepared separately since they are not included in the ECG recording.

Type and specifications of collected data

The manufacturers use different formats for ECG recording. When used as training data, ECG data are often converted to chronological data (text data). For standard 12 lead ECG, the matrix data consist of 12 columns × time points. However, there is a method to capture the ECG waveform as an image. Individual leads or 12 leads may be collected as an aggregate, as image data when using the method.

Current scale of data collection

For prediction of paroxysmal arrhythmia, data were collected from 2,700 patients at 7 institutions, mainly at Tokyo Medical and Dental University and Jichi Medical University, to predict the prevalence of paroxysmal atrial fibrillation. Standard 12 lead ECG with normal sinus rhythm was used at the time of data collection. Annotation data were prepared and linked separately from the ECG text data for cardiac event recording. The generated data were collected with the same model of electrocardiograph at multiple facilities (2 university hospitals, 5 cardiovascular centers) and in different environments (hospital laboratory, patient room). The factors that may affect the results, e.g., use of antiarrhythmic agents, were specifically excluded. Note that the annotation for paroxysmal atrial fibrillation includes recording of actual atrial fibrillation attack to ensure the diagnosis, whereas asymptomatic paroxysmal atrial fibrillation may have been overlooked because ECG records are not available throughout the past. Caution should be used since this problem always arises as a limitation of control data when ECG data are used to predict prevalence/onset of paroxysmal disease.

Type of linked medical information

The information linked to ECG data may include characteristics of the patient, examiner's information (the information should desirably be recorded by a clinical laboratory technician or a physician to ensure the accuracy of ECG electrode placement), device information (evaluation of the features of the electrocardiograph), environmental information (noise outside the laboratory, room temperature, humidity), patient information (annotation data on paroxysmal arrhythmia mentioned earlier), comorbidities and past diseases that may be reflected in the ECG information, use of antiarrhythmic agent, and electrolyte data (blood biochemical test). ECG assessment made by a specialist will be needed when using the data for automatic ECG diagnosis.

Need for data collection in different individuals, regions, medical institutions, countries, and time periods

Although the ECG procedure has been standardized, data should be collected at multiple institutions and in different environments, taking into account the influence of external noise contamination associated with measurement in different environments. Annotation data should be collected after taking into account the regional difference in disease distribution.

Dependence on human procedures

Although the ECG procedure has been standardized, data from ECG performed by unskilled healthcare professionals may have variable quality due to incorrect electrode placement. To ensure proper database creation, personnel performing ECG should preferably be limited to specialists such as clinical laboratory technicians and cardiologists.

Diversity of imaging/recording devices

Among electrocardiographs approved as medical devices, 12 lead electrocardiographs are supposed to have a certain level of data quality. However, the sampling rate and the quantization bit rate may be different depending on the measurement conditions. The quality of Holter ECG signals may be greatly affected by the electrode placement, this tendency is particularly strong in small single-channel electrocardiographs for long-term monitoring, which have been widely used in recent years.

Diversity of devices used

The ECG electrodes include seal-type electrodes using conductive gel, clip-type electrodes, and suction electrodes. Development of shirt electrodes using conductive fiber is in progress. The characteristics of electrodes may affect the quality of ECG signals in the future.

Main annotation method and method of labelled data generation

ECG data will be annotated at the time of collection. If a new arrhythmic episode occurs subsequently, the data on the disease will be newly annotated. At this time, the problem arises as

to whether the relevant ECG should be classified into the group of arrhythmic episodes or annotated as a pre-onset record within the repeated ECG data of the same subject. No standard is available for solving this problem. It is entirely conceivable that the annotation may change depending on the analytical method. Therefore, it is necessary to compile the database including the date and time of ECG recording and onset of arrhythmic episode when the presence or absence of arrhythmic episode is given as annotation data.

Expected use

There are two major usages of a model constructed based on an ECG database: (1) To automatically diagnose the change in the ECG waveform at the time of recording, and (2) to predict the onset of paroxysmal arrhythmia being not present at the time of recording. Annotation is included in the ECG record to provide definite annotation data in usage (1) while annotation data are separated from the ECG record in usage (2).

ECG database issues

There are no global standards for ECG data. Although ECG data are usually converted to text data and collected, temporal resolution (sampling rate), dynamic range, and quantization bit rate vary depending on the device type and the recording conditions. The recording time is 10 seconds with standard 12 lead ECG in general; however, recording time is not standardized.

Collection of ECG data is relatively easy. Data can be obtained from the same subject multiple times. However, data from the same subject does not necessarily reproduce the same data. Long-term temporal changes may include changes in ECG information due to aging or progression of the underlying disease. Short-term changes, e.g., circadian or diurnal variations, may also be reflected in the ECG. Because ECG repeatedly records the waveform associated with the heart rate cycle, features may be included in minute changes in the ECG waveform per heartbeat. These are called time variation (variability), which has different information from static ECG data. Thus, data obtained in different time phases are not completely independent data but still need to be handled as data containing different information. Data obtained in the same time phase need to be handled as data containing short-term time variations.

As mentioned earlier, diagnosis based on lifelong continuous ECG record is needed to provide accurate annotation data on paroxysmal arrhythmic diseases.

7-4. Outline of endoscopic image database

Purpose of database construction

The purpose of database construction is to record gastrointestinal endoscopic data for direct diagnosis. The purpose of database construction may be generated for endoscopic treatment in the future. However, objective evaluation of techniques as in surgery videos is not intended for now. The major purpose is to generalize and enhance gastrointestinal endoscopic diagnosis by enabling accurate diagnoses equivalent to those made by specialists. However, the term diagnosis used in this section does not simply mean determining the disease name and the diagnostic name

or detecting the disease. More complicated assessment such as disease staging and prediction of accidental symptoms/complications is also part of the diagnosis.

Type and specifications of collected data

The target data to be collected in the area of gastrointestinal endoscopy are still images and videos. Static images have already been accumulated in a considerable number of institutions. How to use the sufficient resources is a major issue specific to the area of gastrointestinal endoscopy. The standards of the existing gastrointestinal endoscopic image filing system is complemented by the general standards such as JPEG and bitmap. The deploy operation would be easy. However, images alone are insufficient as a collected data resource as described later. A large quantity of endoscopic images can be interpreted individually. It is important not only to make an initial diagnosis of gastric cancer based on the image, but also to tag and accumulate diagnostic information on an extremely large number of lesions diagnosed with gastrointestinal endoscopy It should be emphasized that tagging with diagnostic names as well as the standardized terms such as the characteristics of disease specified in the Japan Endoscopy Database (JED) project will widen the range of research and operations.

On the other hand, no certain video format has been established. However, the Japan Gastroenterological Endoscopy Society (JGES) has been working on the development of a tool to enable easy tagging of static images clipped from videos and annotation of an enormous amount of generated static images. A highly useful tool should be provided to researchers and developers for gradual standardization rather than mandated standardization.

Current scale of data collection

Although the images accumulated for the support of the AMED or other development purposes differ depending on the disease area, a survey of multiple institutions found 30,000 disease images and an additional 130,000 normal images were collected from gastric cancer patients, 10,000 disease images were collected from patients with inflammatory bowel disease, and 20,000 images were collected to develop an initial algorithm for the recognition of the site of endoscopy in the stomach. An enormous number (540,000) of images of duodenal papilla were accumulated in multiple disease areas. On the other hand, some gastrointestinal endoscopic databases have been commercialized. The commercial databases have been accumulating images separately from the JGES, making it difficult to illustrate the whole picture here. Some gastrointestinal endoscopic databases have been commercialized by companies while some others are under development for commercialization. Not all images have been accumulated in the same manner. Some databases are used only within the theoretical scope of company or only available for operation by researchers, making it very difficult to develop a general, integrated database. One hundred thousand still colonoscopic images have been clipped from videos and provided for research purposes.

Type of linked medical information

Regardless of static images or videos, use of linked data may be considered for AI learning based on static images. It is important to tag detailed information according to the JED data terms standardized by the JGES as mentioned earlier. In gastrointestinal endoscopy, the organ and the site of the image is also important information. Location information is an important aspect especially for automatic site recognition. It is also important to add items according to the purpose of research and development. The tool to clip static images from videos described earlier also accommodates addition of addable tag information.

The JED data include qualitative diagnosis (disease name), affected site, size, and existing findings (information on endoscopic findings) as primary diagnostic information. Standardized terms are also available for characteristics of the patient including family history, diseases history, and preference. Data are stored in conformity to the JED format to provide organized information. In addition, the endoscope (scope) model and the imaging conditions have also been standardized as particular information on gastrointestinal endoscopy.

Need for data collection in different individuals, regions, medical institutions, countries, and time periods

Some diseases are more prevalent in specific regions. It is the same in Japan and other countries. In addition, heterochronic information has so many meanings. Temporal change is entirely possible even in the same patient. Heterochronic data should be handled separately even in the same individual. The concept of normal is very important in research and development. It should be emphasized that it is necessary to accumulate normal data as normal not as disease without a name.

Another point is at least about 40 static images are stored in one gastrointestinal endoscopy examination. More than 100 images are accumulated in one examination for a rare disease. However, the major characteristic of gastrointestinal endoscopy is that accumulated images include no-lesion images even if lesions are found in a single examination. Currently, tagging is limited to Information for individual examination, however, individual tag information is needed for all images to be used for ML as with the application for clipping static images from videos. This is also a major characteristic of gastrointestinal endoscopy. This aspect is significantly different from other disease areas where only the lesion images are used specifically. A large number of images will be easily obtained with detailed tagging. How to take advantage of this feature is a future challenge.

Dependence on human procedures

Certain stratification will be important because the quality of endoscopic/ultrasound images heavily depends on the skills and experience of the clinician performing the procedure compared with highly objective images such as radiological and microscopic images. Years of experience is also a reasonable factor as for surgery videos. How to handle information of patients and physicians is a major theme and a future issue.

Diversity of imaging/recording devices

The name of the endoscope (scope) and the information on the various conditions for special light observation called optical digital method are important. However, frequent change of the conditions for special light observation during an examination session is characteristic of the device. In this regard, the conditions should ideally be added to individual images, not to individual examinations, as tag information. The development of such a function is progressing, albeit gradually.

Diversity of devices used

A number of different devices are used for endoscopy alone. We are also working to make the devices recognizable.

Main annotation method and method of labelled data generation

Although it may depend on the image processing and the learning methodology applied to the AI research, it is unlikely that only the lesion is captured in the image. The lesion is generally shown in part of the image. Therefore, annotation to accurately trace the lesion is important for lesion recognition.

Expected use

Expected purposes of use of the database include: (1) Detection, (2) differential diagnosis, (3) deviation (check to see if the test was performed according to specified rules), (4) staging (including effects of therapy evaluation), and (5) prediction (difficulty level of the procedure, incidence of accidental symptoms/complications).

Endoscopic image database issues

Ethical issues are the greatest challenges to overcome. Use of images originally collected for research purposes for other purposes is prohibited. The database is fundamentally not designed for general use. A database with a high operational flexibility should be created after obtaining individual consent.

7-5. Common database challenges and issues

Large data volume

A gigantic storage capacity is required. This is a common challenge to all databases. For example, a capacity of about 18 GB is required for one FHD video of 3-hour surgery. Seventy-two TB will be required to store 4000 videos. The required storage capacity will more than double if data are backed up before standardizing the videos and processed while protecting personal information. The disadvantages of having a large storage capacity include time-consuming training and testing and a considerable expense for cloud management.

Complexity of creating an annotation dataset

Annotation is quite a specialized and complicated process requiring diagnoses and judgment of physicians. For example, the annotation tool developed by the JGES to contribute to gastrointestinal endoscopic research merely reduces the already immense amount of time.

Defrayment of various expenses related to the database

Defrayment of expenses for creation, operation, and management of a database is a major challenge. The AMED research fund is used to collect data for many of the databases in Japan. The maintenance and operational costs will not be covered once the research project is completed. In this sense, one of the major challenges is to commercialize the database for continuous data collection without relying on research funds in the future.

The database of digital pathological images was initially planned to be stored in a secure cloud storage provided by a company. Since the annual estimate was several tens of million yen, however, the database is currently stored in a cloud storage provided by a third party, the annual cost of which is several million yen. The JGES members are responsible for the database operation. The Research Committee of the JSP allows its members to use the images for human lifelong learning and image analysis for AI development for the purpose of basic academic research at no cost after reviewing the study protocols. The cloud operation cost will be generated by offering unannotated WSIs for 500 yen per image to company-funded research projects, educational projects taken on by companies, or company-initiated AI development. Regarding the surgical video database, a company was established to continue the maintenance and operation of the database after the completion of the AMED project.

8. Discussion on data for the development of SaMD using ML/DL (training data, validation data, test data)

There are three types of data used for the development of SaMD: Training data, validation data, and test data. For example, training data are used for the weight parameter in learning process of a DL model, and validation data are used to determine hyperparameters such as the structure and the number of training epochs of a DL model. In a broad sense, these two types of data are categorized as data for training. On the other hand, test data are used to estimate the performance of a trained SaMD when applied to real world data after marketing.

The three types of data are primarily used in the model development phase such as in the process of cross validation. When considering the conditions required for the data, which of the three types of data are being discussed should be made clear. Test data are also used in the approval and certification review process for SaMD as a medical device to evaluate whether the performance of the SaMD is adequate for the expected healthcare application. Note that the conditions and assumptions mentioned in the following discussion apply more strictly to test data for approval and certification review compared with test data used in development.

In the ML area, it is often assumed that the data for training (including training data and validation data) is independent and identically distributed to the test data, i.e., the data are assumed to have been independently sampled from the real world data to be applied and have an identical distribution. The assumption that training data and test data are independent and have identically distribution is discussed below.

Independence

Needless to say, the independence of the training /validation data used for development and the test data used for performance evaluation of the developed DL model must be independent from each other to maintain the validity of the test. In the assumption of independence between training data and test data, use of completely identical data for both training and testing is prohibited. Any actions that raise questions about data independence, such as intentionally adding images similar to the test data to the training data, should also be strictly prohibited. However, it is necessary to sufficiently discuss whether the idea should be easily extended to uniformly prohibit using the data of the same subject as both training and test data.

This is an important issue also discussed in the aforementioned cases of database development in the Chapter 7. Overly restricting the use of data from the same subject based on the emphasis of data independence may result in inefficient access to valuable medical data. For example, the independence of data of the same subject may exceptionally be ensured if the characteristics of the data greatly differ due to different types of diseases, different parts of human body, and separate timing of data acquisition. It is necessary to allow for conditional options after careful evaluation of the data independence of the same subject based on the type of disease, the disease site, and the timing of imaging. When relaxing the conditions, it is essential to explain the validity scientifically and objectively to keep the independence between the training data and the test data

unquestionable. To accurately estimate the performance of a developed model, it is important to maintain the independence of the test data from the data used for training and validation during the development.

Identity of the distribution of training data and test data

Independence identical distribution between data for training (including training data and validation data) and test data is related to the discussion of Bias in Chapter 3. The database creation project in various areas described in Chapter 7 will provide data to reasonably demonstrate the identity if the application area is specified. In the future development of ML data in new areas and databases to be provided, the database specifications should desirably be determined from the perspective of these discussions.

Regarding the test data used to evaluate trained performance of a DL model, it should be assumed independent of the training data sampled from the real world distribution to be applied to the SaMD. On the other hand, in a training phase, a different discussion is needed a whether the assumption that the training data are sampled from the identical real world distribution should be strictly applied. According to some recent reports, concurrent use of training data that do not meet the condition for identity may improve the ML performance in the training phase. For example, there are numerous reports of application of a model trained with a large number of natural images to medical images [44, 45]. In addition, models trained with an incommensurable quantity of data (e.g., natural images other than medical images, text information) called foundation models [46] are now available. An example is a highly versatile model designed with more than 10^7 images or more than 10^9 segmentation masks and applied to a variety of types of images, such as a Segment Anything Model (SAM) [47]. Cases have now been reported in which the assumption that the training data has the identical distribution as the test data does not hold. For example, for image classification, superior performance is achieved by first training on a large amount of non-medical (natural) data and then on medical image data. Most SaMD without DL were designed using training data independently sampled from the identical real world distribution in the target area of application. This has been considered as a normal condition in the development of DL models used in healthcare based on databases in various clinical areas mentioned earlier. However, the circumstances are changing now. The assumption of identical distribution of training data and test data as in models trained with non-medical images may not hold true in many future cases. Therefore, it may not be appropriate to strongly insist on the assumption that training and validation data used for the development have the same distribution as the real world data of the target area of application in a future SaMD review.

AI technologies including DL are evolving to be highly flexible, e.g., the emergence of techniques to accommodate different distributions from those of training data, such as transfer learning [48] and domain generalization [49]. There is no need to be excessively obsessed with assumption that the training and validation data used in development have the identical distribution as the real world data in the intended application area. The more important question is whether or not the desired performance of SaMD with application of the real world data can be

appropriately estimated using the prepared test data. To this end, it should be emphasized again that it will be more important to assume the test data have been sampled from the real world data to be applied to the SaMD independently of the training data (including validation data) and scheme to prevent leakage of the test data information to the developer when reusing the test data pointed out in Chapter 4.

Underspecification of DL

This final section discusses the issue of underspecification [50] of the DL model, which has arisen recently.

As mentioned in the beginning of this Chapter 8, test data are used to estimate the performance of trained SaMD when post-market real world data are applied. To accurately estimate the performance, it is important that the test data and the real world data have identical to.

In reality, however, there are multiple models whose performance on post-market real world data differs, but whose performance is indistinguishable in the test data at hand on the development side. Therefore, it has been reported that there are cases where multiple models that showed the same performance during development show different performance when applied to actual real world data, and this is positioned as a problem known as underspecification. The performance of a SaMD with this problem may be significantly decreased compared with the performance at the time of approval when post-market data are applied. A method called a stress test, in which the test data are intentionally perturbed to create a variation around the data, has been proposed as an evaluation method for underspecification [51]. However, research on how to address underspecification is still ongoing as an unsolved issue, and there are still some points where the evaluation method has not been finalized.

The research trend of underspecification should be continuously monitored.

9. Summary

This report discussed the challenges in development of ML-based SaMD and its implementation in society and summarized what data-related topics are to be considered.

Various considerations are required in terms of "bias." The data used for training, validation, and test must be homogeneous with the statistical properties of the target patient population for the intended use of SaMD. This report also pointed out the unconscious user bias and potential human bias in data selection and annotation, to which not much attention has been hitherto paid. These concepts are common to the concept of "Good Machine Learning Practice" recommended by the regulatory authorities of the respective countries.

One of the features of ML-based SaMD is that it can be designed to change its performance by accumulating real world data after marketing and through continuous training. The FDA has been testing the Pre-Cert to address the issue of how to evaluate such products. Many issues have been pointed out to date. This indicates a difficulty in scientifically estimating post-market performance before marketing even if a method of post-market learning is clearly defined. Note that a bias may occur when selecting data to be used for re-training based on the results of real world performance evaluation. This report presented current scientific knowledge about this issue, including method development research to avoid the issue.

Difficulty of collecting a large quantity of training, validation, and test data of a certain quality has been pointed out as one of the difficulties in ML application in the medical field. Training based on a large quantity of data collected for other published purposes is also often implemented in the medical field. This report pointed out that a bias unrecognizable to humans may affect the performance of AI models developed as training data, possibly causing a dissociation between the performance in the developmental phase and the performance with real world data, based on the past studies. In addition, this report discussed the use of numerical simulation data as a new method to obtain training data and identified related issues. The concepts presented in the "Report of the Subcommittee on Computer Simulation" [3] published in the past by the PMDA Science Board, such as the validation of the numerical calculation model, can be used. With the recent progress in AI image generating technology, application of image generating technology to compensate for a lack of training data will also be a future research topic. The (statistical) properties of data used for training, validation, and evaluation will need to be homogeneous with that of the target patient population for the intended use of SaMD.

This report also pointed out that many of randomized clinical trials published as scientific reports are not prospective in design and involving risk of bias. To address the issue of how to conduct scientifically appropriate evaluations using limited data in the medical field where collecting a large quantity of test data of a certain quality is difficult, rational consideration will be required on a case-by-case basis while referring to many precautions pointed out in this report based on conventional medical technology assessment.

The current status of medical database construction for ML application was investigated in medical image areas, and common issues were extracted based on the discussions at the

Subcommittee. The precautions pointed out in this report are useful for enhancement of medical database development and design and operations of new databases creation in the future.

There are three type of data in ML: training data used to train ML models, so-called labelled data, validation data used to verify the model performance obtained through the training process and also used to improve the model during the development process, and test data used in the course of development, test data used to evaluate the performance of completed ML models for medical device approval or certification. The related issues are summarized in Chapter 8 "Discussion on data for the development of SaMD using ML/DL (training data, validation data, test data)." There is tendency to consider collectively and require the same quality for these three types of data. However, it is necessary to define quality requirements while keeping in mind the purpose of data usage, e.g., idea of possible use of simulation data as training data with appropriate bias control, and the existence of knowledge that training with data not homogeneous to the data obtained in the intended field of application (e.g., training with a large quantity of natural image data for medical image processing) is effective in achieving superior performance as a result. Understandably, however, it is currently unacceptable to use simulation data as test data for performance evaluation of ML-based (e.g., DL) SaMD, unless it is scientifically shown that the simulation data appropriately reproduce the real phenomena in line with the purpose to be evaluated in the test. The current stance of this report is that test data should be assumed to be sampled from the same distribution of real world data independently of training data (including validation data used during the development). Considering the characteristic difficulty of obtaining a large quantity of high-quality medical data, it is necessary to further discuss the requirements for the data in the future.

The following points must be noted when using the data recorded in the created database as approval application data for the developed SaMD:

- Documents required for the review process, including compliance inspection, should be prepared for submission in advance with due considerations to the related laws and regulations such as the Personal Information Protection Act and the Clinical Trials Act. Submission and discussion of sufficient data for the review may not be possible if the access to the necessary data is denied due to violation of the related laws and regulations. Careful attention must be paid to this point so as not to make the review process difficult.
- Note that the data need to be anonymized in an appropriate manner after obtaining the patient's consent to the purpose of data use (for product development including approval application, or for a commercial purpose) when using the collected medical images and the related clinical information.

Expectations for the introduction of ML-based AI in the medical field are growing. Appropriate design of the development process and scientific and rational performance evaluation with reference to the precautions presented in this report will lead to the social implementation of safe and effective DL-based AI medical devices.

Reference

[1]. The Science Board. Issues and Proposals on AI-based Medical Devices and Systems 2017 (2017) https://www.pmda.go.jp/files/000224080.pdf (in Japanese)(accessed 2023-07-26)

[2]. Ministry of Health, Labour and Welfare. Partial Revision of the Guideline of determining whether software is classified as a medical device. Notification No.0331-1 of PSEHB/MDED, Notification No.0331-4 of PSEHB/CND (March 31, 2023) https://www.mhlw.go.jp/content/11120000/001082227.pdf (in Japanese)(accessed 2023-07-26)

[3]. The Science Board. Report on Reviewing Medical Device Software Using Computer Simulation. (2021) https://www.pmda.go.jp/files/000240657.pdf (in Japanese)(accessed 2023-07-26)

[4]. U.S. Food and Drug Administration. Discussion Paper and Request for Feedback, "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)." (2019) https://www.fda.gov/media/122535/download (accessed 2023-07-26)

[5]. U.S. Food and Drug Administration. Draft Guidance for Industry and Food and Drug Administration Staff, "Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML)-Enabled Device Software Functions." (2023) https://www.fda.gov/media/166704/download (accessed 2023-07-26)

[6]. U.S. Food and Drug Administration. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. (2021) https://www.fda.gov/media/145022/download (accessed 2023-07-26)

[7]. U.S. Food and Drug Administration. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Device. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices (accessed 2023-07-26)

[8]. U.S. Food and Drug Administration. The Software Precertification (Pre-Cert) Pilot Program: Tailored Total Product Lifecycle Approaches and Key Findings. (2022) https://www.fda.gov/media/161815/download (accessed 2023-07-26)

[9]. Division of Medical Devices, National Institute of Health Science. Report on the 2021 Survey Project on Advanced Medical Devices including AI and Mobile Applications. (2022) https://dmd.nihs.go.jp/samd/R3_report.pdf (in Japanese)(accessed 2023-07-26)

[10]. U.S. Food and Drug Administration. Health Canada, and UK Medicines and Healthcare products Regulatory Agency. Good Machine Learning Practice for Medical Device Development: Guiding Principles. (2021) https://www.fda.gov/media/153486/download (accessed 2023-07-26)

[11]. European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence

Act) and Amending Certain Union Legislative Acts.(2021) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206 (accessed 2023-07-26)

[12]. Ministry of Health, Labour and Welfare. Release of Evaluation Indices for Next-Generation Medical Devices　Exhibit 4　[Evaluation Indices for medical image diagnosis support systems using artificial intelligence technology].Notification No.0523-2 of PSEHB/MDED (May 23, 2019) https://dmd.nihs.go.jp/jisedai/tsuuchi/%E8%96%AC%E7%94%9F%E6%A9%9F%E5%AF%A9%E7%99%BA0523%E7%AC%AC2%E5%8F%B7_%E5%88%A5%E7%B4%994.pdf　(in Japanese)(accessed 2023-07-26)

[13]. Pharmaceuticals and Medical Devices Agency. Points of Review for Software as a Medical Device (SaMD). https://www.pmda.go.jp/review-services/drug-reviews/about-reviews/devices/0047.html (in Japanese)(accessed 2023-07-26)

[14]. Ministry of Health, Labour and Welfare. Publication of Guidance for Appropriate and Prompt Approval and Development Based on the Characteristics of Programmed Medical Devices. (May 29, 2023) https://www.mhlw.go.jp/hourei/doc/tsuchi/T230530I0080.pdf (in Japanese)(accessed 2023-07-26)

[15]. IMDRF/AIMD WG/N67. Machine Learning-enabled Medical Devices: Key terms and Definitions. (2022) https://www.imdrf.org/sites/default/files/2022-05/IMDRF%20AIMD%20WG%20Final%20Document%20N67.pdf (accessed 2023-07-26)

[16]. Hall M et al. A Systematic Study of Bias Amplification. arXiv. 2022;2201.11706. (doi: 10.48550/arXiv.2201.11706)

[17]. Bercean BA et al. Evidence of a cognitive bias in the quantification of COVID-19 with CT: an artificial intelligence randomised clinical trial. Scientific Reports.2023;13.4887. (doi: 10.1038/s41598-023-31910-3)

[18]. Heave WD.　MIT Technology Review. Hundreds of AI tools have been built to catch covid. None of them helped. (2021) https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/ (accessed 2023-07-26)

[19]. Pianykh OS et al. Continuous learning AI in radiology: implementation principles and early applications. Radiology. 2020;297(1):6-14. (doi:10.1148/radiol.2020200038)

[20]. Subbaswamy A et al. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics. 2020;21(2):345-352. (doi:10.1093/biostatistics/kxz041)

[21]. Ministry of Health, Labour and Welfare. Handling of post-approval change management protocol (PACMP) of medical device. Notification No.0831-14 of PSEHB/MDED (August 31, 2020). https://www.pmda.go.jp/files/000236900.pdf　(in Japanese)(accessed 2023-7-26)

[22]. Shimada K et al. Simulation of Postmarket Fine-tuning of a Computer-aided Detection System for Bone Scintigrams and Its Performance analysis. Advanced Biomedical Engineering. 2023;12:51-63. (doi:10.14326/abe.12.51)

[23]. Lange MD et al. A Continual Learning Survey: Defying Forgetting in Classification Tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022;44(7):3366-3385. (doi: 10.1109/TPAMI.2021.3057446)

[24]. Yoshimura I. Comments on Design Considerations in Controlled Clinical Trials for Confirmatory Purposes - For Better Understanding of "Statistical Principles for Clinical Trials"-. (in Japanese) Proceedings of the Institute of Statistical Mathematics. 1998; 46(1):81-95.

[25]. Dwork C et al. The reusable holdout: Preserving validity in adaptive data analysis. Science. 2015;349(6248):636-638. (doi:10.1126/science.aaa9375)

[26]. Freedman DA. A note on screening regression equations. The American Statistician.1983; 37(2):152–155. (doi:10.2307/2685877)

[27]. Sakuma J. [Privacy preservation in data analytics] (in Japanese) Kodansha. 2016

[28]. Gossmann A et al. Test Data Reuse for the Evaluation of Continuously Evolving Classification Algorithms Using the Area under the Receiver Operating Characteristic Curve. SIAM J. MATH. DATA SCI.2021;3(2):692-714 (doi:10.1137/20M1333110)

[29]. Roelofs R et al. A Meta-Analysis of Overfitting in Machine Learning. Neural Information Processing Systems. 2019. (https://api.semanticscholar.org/CorpusID:207979247)

[30]. Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006.

[31]. Medical Device Evaluation Division, Pharmaceutical Safety and Environmental Health Bureau, Ministry of Health, Labour and Welfare. Report on the deliberation results.(2022) https://www.pmda.go.jp/medical_devices/2022/M20220516002/112714000_30400BZX00101_A100_4.pdf (in Japanese)(accessed 2023-07-26)

[32]. Nagendran M et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020; 368 :m689 (doi:10.1136/bmj.m689)

[33]. Shimron E et al. Implicit data crimes: Machine learning bias arising from misuse of public data. Proc Natl Acad Sci U S A. 2022;119(13):e2117203119. (doi:10.1073/pnas.2117203119)

[34]. Nguyen DP et al. Reinforcement learning coupled with finite element modeling for facial motion learning. Computer Methods and Programs in Biomedicine. 2022; 221:106904. (doi:10.1016/j.cmpb.2022.106904)

[35]. Nguyen PCH et.al. Synthesizing controlled microstructures of porous media using generative adversarial networks and reinforcement learning. Scientific reports. 2022;12(1):9034. (doi:10.1038/s41598-022-12845-7)

[36]. Thambawita V et.al. SinGAN-Seg: Synthetic training data generation for medical image segmentation. PLOS One. 2022;17(5):e0267976. (doi:10.1371/journal.pone.0267976)

[37]. Viceconti M et.al. In silico trials: Verification, validation and uncertainty quantification of T predictive models used in the regulatory evaluation of biomedical products. Methods. 2021; 185:120-7. (doi:10.1016/j.ymeth.2020.01.011)

[38]. Sasaki T. The Japanese Society of Pathology JP-AID and Pathology Diagnostic Artificial Intelligence Development. (in Japanese) Byori to rinsho(Pathology and Clinical Medicine). 2017;35(11) 1058-1061.

[39]. Uozaki H et al. 7th Development of Digital Image Collection Infrastructure at the Japanese Society of Pathology. (in Japanese) Byori to rinsho(Pathology and Clinical Medicine). 2018;36(10):1017-1021.

[40]. The Japanese Society of Pathology, Japanese Society of Digital Pathology. Digital Pathology System Technical Standards for Pathological Diagnosis (3rd Edition). (2018) https://pathology.or.jp/news/pdf/kijjun-181222.pdf (in Japanese)(accessed 2023-07-26)

[41]. Sasaki T. Development of AI Algorithm to Support Pathological Diagnosis: Initiatives of the Japanese Society of Pathology. (in Japanese) Iryou kikigaku (The Japanese Journal of Medical Instrumentation). 2019;89(6):526-532.

[42]. Sasaki T. Challenges and Prospects for AI Programs in the Pathological Diagnosis Field. (in Japanese) Modern Media. 2022;68(3):74-80.

[43]. Sasaki T. Pathological Image Diagnosis by AI. (in Japanese) Bone Joint Nerve. 2021;11(2):227-233.

[44]. Litjens GJS et al. A survey on deep learning in medical image analysis. Medical Image Analysis. 2017; 42:60-88. (doi:10.1016/j.media.2017.07.005)

[45]. Suganyadevi S et al. A review on deep learning in medical image analysis. International Journal of Multimedia Information Retrieval. 2022;11(1):19-38. (doi:10.1007/s13735-021-00218-1)

[46]. Bommasani R et al. On the Opportunities and Risks of Foundation Models. arXiv. 2022. (doi:10.48550/arXiv.2108.07258)

[47]. Kirillov A et al. Segment Anything. arXiv. 2023. (doi:10.48550/arXiv.2304.02643)

[48]. Fuchao Y et al. A Survey on Deep Transfer Learning and Beyond. Mathematics. 2022;10(19):3619. (doi:10.3390/math10193619)

[49]. Zhou K et al. Domain Generalization: A Survey. IEEE Trans. on Pattern Analysis and Machine Intelligence. 2023; 45(4):4396-4415. (doi:10.1109/TPAMI.2022.3195549)

[50]. D'Amour A et al. Underspecification Presents Challenges for Credibility in Modern Machine Learning. Journal of Machine Learning Research. 2022;23(226): 1-61. (https://jmlr.org/papers/volume23/20-1335/20-1335.pdf) (accessed 2023-07-26)

[51]. Eche T et al. Toward Generalizability in the Deployment of Artificial Intelligence in Radiology: Role of Computation Stress Testing to Overcome Underspecification. Radiology: Artificial Intelligence 2021;3(6) (doi:10.1148/ryai.2021210097)

Member List of the Subcommittee on Software as a Medical Device Utilizing AI and Machine Learning

○ITO Masaaki
Deputy Director・Head of Department of Colorectal Surgery・Head of the medical device development promotion division, National Cancer Center Hospital East

◎SAKUMA Ichiro
Professor, Medical Device Development and Regulation Research Center, School of Engineering, The University of Tokyo

SASAKI Takeshi
Project Professor, Next-Generation Pathology Information Networking, Graduate School of Medicine, The University of Tokyo

SASANO Tetsuo
Professor, Director of the Department of Cardiovascular Medicine, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University (TMDU)

SAWA Tomohiro
Professor, Medical Information System Research Center, Teikyo University

SHIMIZU Akinobu
Professor, Institute of Engineering, Tokyo University of Agriculture and Technology

JINZAKI Masahiro
Professor, Department of Diagnostic Radiology, Keio University School of Medicine

TAKEDA Toshihiro
Professor, Department of Integrated Medicine, Medical Informatics, Osaka University Graduate School of Medicine

TANAKA Kiyohito
General Manager of Internal Medicine/ Medical Information Office Manager, Japanese Red Cross Kyoto Daini Hospital

CHINZEI Kiyoyuki
Prime Senior Researcher, Health and Medical Research Institute, National Institute of Advanced Industrial Science and Technology

TONOMURA Keiji
Lawyer, Nagashima Ohno & Tsunematsu

○NAKAOKA Ryusuke
Head of Implantable Devices Section, Division of Medical Devices, National Institute of Health Sciences

NAKATA Norio
Professor, Recent Technical Development of Artificial Intelligence for Medical Research, Research Center for Medical Sciences, The Jikei University School of Medicine

NAKADA Haruka
Chief of COI Management Section, Bioethics Division, Center for Research Administration and Support, National Cancer Center Japan

MURAGAKI Yoshihiro
Professor, Manager of Center for Advanced Medical Engineering Research and Development（CAMED）/ Graduate School of Medicine, Kobe University

MORI Kensaku
Professor, Graduate School of Informatics, Nagoya University
Director, Information Technology Center, Nagoya University

YOKOI Hideto
Professor, Department of Medical Informatics, Kagawa University Hospital

◎Chairperson ○Vice Chairperson