Molecular Imaging CRO Network

Micron's ViewPoint

Clinical Evaluation of Intelligence Decision Support Systems for Diagnosis Based on Medical Imaging: The Importance of the Reference Standards

- For the Social Implementation of Medical AI -

Micron Inc.

Contents

Micron Inc.

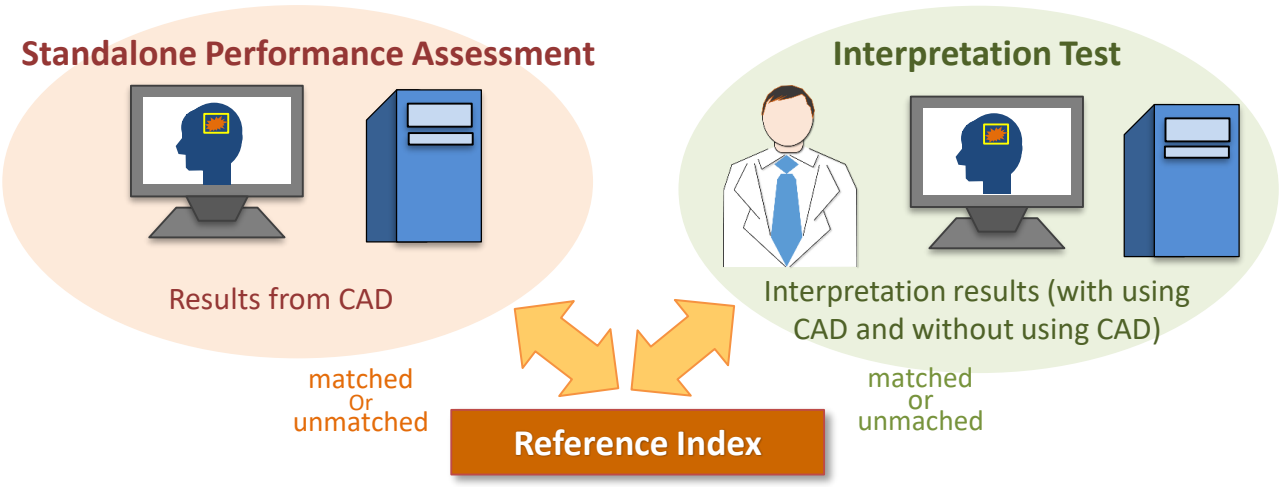
Introduction

CAD (Computer-Aided Detection/Diagnosis), which enables the medical image quantitative analysis on a computer, and detects and diagnoses the presence or absence of diseases and conditions, is a computer e-decision support software program for diagnosis that supports a doctor's diagnosis as "a second opinion". Recently, CAD with artificial intelligence (AI) (hereinafter called "CAD with AI") has been increasingly implemented in clinical settings. In the United States, over 30 CAD with AI programs have been approved by FDA. Furthermore, in Japan, endoscopic image diagnostic support software, which is also CAD with AI, received its first approval last year. As described above, CAD with AI, which assists a doctor's diagnosis by using medical images such as radiographic images, endoscopic images, and ultrasonic images, has been actively developed both in Japan and abroad. The era of its implementation in clinical practice has arrived.

Objective diagnostic performance is one of the determinative factors for the introduction of intelligence decision support systems for diagnosis based on medical imaging (medical image diagnosis support system) that supports physicians to make their diagnosis in clinical practice. In addition, this is an important application data point for obtaining the certification or approval as a medical device. Diagnostic performance of the medical image diagnosis support system is evaluated at the stand alone assessment to evaluate the system's performance, and the interpretation test to evaluate how much the readers would be influenced by the system. Sensitivity, Specificity, or ROC (Receiver Operating Characteristic), and AUC (Area Under the Curve) are common indices for evaluating the diagnostic performance of these tests. These evaluation indices are calculated based on the diagnostic method with the highest degree of accuracy at present. The diagnostic method to be used as a reference is called the reference standard*. The reference standard is evidence to demonstrate either "the subject definitely has the target disease of this study" or "the subject definitely does not have the target disease." Low accuracy of the reference standard may lead to the low reliability of the interpretation test.

This document describes an overview of the reference standard determinations for an interpretation test of the medical image diagnosis support system.

Figure 1: Standalone Performance Assessment and Interpretation Test



*Reference Standard is also called as the Gold Standard or the Standard of truth.

Reference #1, 2, 3, 4, 5 and 6

Micron Inc.

Performance Assessment Index for Medical Image Diagnosis Support Systems

Just observing detection and diagnosis abilities of the medical image diagnostic support system alone is often not sufficient to evaluate the usability of the system. This means that it is important to prove how useful a medical image diagnosis support system is in real clinical practice.

The clinical usefulness of the medical image diagnosis support system can be proved by performing an interpretation test by readers such as physicians or medical workers. An interpretation test evaluates the diagnostic accuracy by visual evaluation, the ability to shorten the interpretation time, and the effect on fatigue. In this document, we focus on the results obtained by visual evaluation (Table 1).

Table 1: Performance Assessment Index

Index	Description
Sensitivity	Percentage of positive diagnostic result (true positive) if the reference standard is positive
Specificity	Percentage of negative diagnostic result (true negative) if the reference standard is negative
Positive Predictive Value	Percentage that the reference standard is positive (true positive) if the diagnostic result is positive
Negative Predictive Value	Percentage that the reference standard is negative (true negative) if the diagnostic result is negative
Positive Likelihood Ratio	Ratio indicating how many times a true positive is more likely to be positive than a true negative
Negative Likelihood Ratio	Ratio indicating how many times a true positive is more likely to be negative than a true negative
AUROC Area under Receiver Operating Characteristic	Area of the curve obtained by plotting all possible values of the threshold on the vertical axis as the true positive rate (sensitivity) and on the horizontal axis as the false positive rate (1 - Specificity).

Reference #5 and 6

Micron Inc.

Performance Assessment Index for the Medical Image Diagnostic Support System

In order to calculate the assessment indices of diagnostic performance such as sensitivity and specificity, it is fundamental to prepare a 2 × 2 quadruple table as follows.

Table 2: Fourfold Table

	Reference Standard Positive	Reference Standard Negative	
Index Assessment Positive	**True positive** a	**False positive** b	**Positive predictive value** =a/(a+b)
Index Assessment Negative	**False negative** c	**True negative** d	**Negative predictive value** =d/(c+d)
	Sensitivity =a/(a+c)	**Specificity** =d/(b+d)	

A distribution map (Figure 2) can be created in accordance with the table above. Threshold* may be adjusted appropriately. By plotting the true and false positive rates, some curves can be delineated. These are the ROC curves (Figure 3). In particular, AUROC has been widely used and regarded as a reliable evaluation index in image interpretation test. Types and methodologies of ROC interpretation test will be detailed explained in our following white papers.

As shown in the table above, calculations of sensitivity, specificity, and evaluation indices, are being discussed on the premise that the reference standard represents "the true patient condition". Such an accurate reference is one of the factors that determine the success of the interpretation test.

Figure 2 Binomial Distribution Map

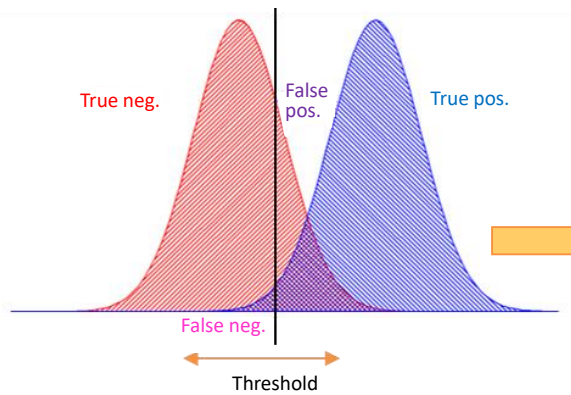
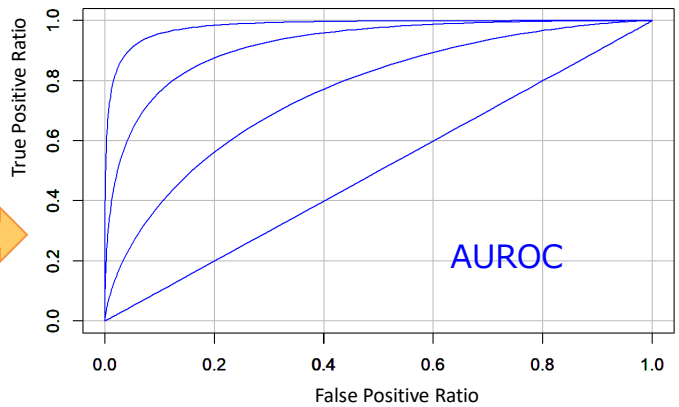


Figure 3 ROC Curves



*Thresholds are often replaced by confidence for the medical image diagnosis support system .

Reference #7

Micron Inc.

Method for Determining the Reference Standard by the Expert Panel

Reference standards shall be determined in the best and most appropriate manner practicable, and the method for determining reference standards shall be carefully designed. The following methods can be used to determine the reference standards:

- Output results obtained from other medical image diagnosis support system
- Already established clinical diagnosis (Example; pathological findings, clinical test results)
- Follow-up results from imaging exams
- Follow-up results other than imaging exams
- Interpretation of expert panels*

Reference standard may be determined by the independent expert panel's interpretation in case it's impossible to obtain pathological finding reports or clinical test results since the interpretation test is a such as retrospective study, or if there is no way to determine reference criteria with clear criteria.

An expert panel which determines reference standard for an interpretation test is structured in accordance with four elements below:

* Group of people who have specialized abilities and knowledge needed for the area.

1. Structure Members

Expert panels are often structured with more than one member because the results derived from a multi-member consensus are likely to lead to a more accurate interpretation results than those of a single member. In addition, when a majority voting style is applied, it is desirable to form an expert panel with an odd number of people to avoid a tie vote.

The clinical backgrounds of the experts is also an important element. Whenever possible, it is desirable to have experts from all fields related to the target disease. In addition, the level of expertise and years of experience in the target disease are also factors when selecting experts.

Reference #4, 5 and 8

Micron Inc.

Method for Determining the Reference Standard by the Expert Panel

2. Information Provided to the Expert Panel

Information provided to the expert panel is determined depending on the interpretation test design. The reasons for which information should be provided to the expert panel, including relevant guidelines, systematic reviews, and key papers should be identified. For example, in addition to imaging information, the patient's medical history, complications, treatment history, physical findings, and blood test results should also be considered.

On the other hand, blinding is also important to avoid bias by clinical information. There are several types of blinding: completely blinded, partially blinded, stepwise blinded, and unblinded (open label). The appropriate blinding type should be chosen based on study characteristics.

3. Determination Process

The reference standards consist of the presence or absence of the target disease, categorical classification of severity and diagnostic accuracy, location of the lesion, etc. The determination process of them is basically multiple independent reviewers' voting. In general, a consensus meeting is not recommended since it may influence each reviewer's decision-making. Even if holding a consensus meeting cannot be avoided, each reviewer should perform the independent review before and determine the diagnostic results. This individual evaluation can also be used to determine the subgroups that do not need to be discussed by the expert panel as a whole. The exclusion of some cases from the expert panel's review may lead reduction of committee members' workloads.

In case of review result discrepancies between each independent reviewer, selecting an additional new expert, or provision of additional information, is one method to determine the review result. Either way, the review result determination processes for the case of each reviewer's review result's discrepancy should be finalized prior to study initiation.

4. Validity of the Expert Panel

It is recommended that review results by the expert panel have reproducibility, because interpretation of the results may be subjective, and inaccuracy in case the reference standard doesn't have a clear decision criteria. To avoid variability of reviewers' review results, objective evaluation criteria should be prepared. Reviewer training prior to the study initiation is also recommended as needed. Besides, quantifying the reproducibility is one of the methods to ensure the reliability of the interpretation test. Concretely speaking, the kappa coefficient for categorization data and the Intraclass Correlation Coefficient (ICC) for continuous data are widely used.

One point to keep in mind in organizing the expert panel's interpretation test is spending a lot of time and workloads. Setting an efficient and optimum design prior to the interpretation test will affect the shortening of the study period. There is also a risk of bias in the interpretation test as in other clinical trials, particularly if it's a retrospective study. If incorporation bias distort the interpretation test results, it may lead to misinterpretation and the appropriate evaluation results may not be obtained. It is important to develop a study design that eliminates bias. The types of biases in the interpretation test, and the countermeasures, will be described in detail in following white papers.

Reference #5, 7, 8 and 9

Micron Inc.

Conclusion

To realize Precision Medicine which delivers an optimal treatment for each patient, more accurate diagnostic exams are in development. In particular, image diagnosis using AI is one of the anticipated prioritized areas in the third AI boom focusing on deep learning.

In order to evaluate the diagnostic accuracy of intelligence decision support systems for diagnosis based on medical imaging in an appropriate manner, it is important to establish suitable performance assessments. If the data from these tests are used for marketing approval application, it is desirable to develop more rational and appropriate test designs.

As a leading imaging CRO (imaging core-lab) from APAC, Micron has been supporting multiple imaging clinical trials. As a clinical research organization for imaging clinical trials, we always propose the optimal design of the clinical trials as a reliable partner of the Sponsor. Please contact us if you're interested in the application and registration of regulatory approval (e.g. Shonin/Ninsho) for "CAD with AI" and intelligence decision support systems for diagnosis based on medical imaging.

Company Overview

Headquarters (Tokyo)	1-3-5 Nihonbashi Chuo Tokyo Phone: +81-3-6262-2830, FAX: +81-3-6262-2831
Osaka Branch	4-5-36 Miyahara Yodogawa Osaka Phone: +81-6-6399-0007, FAX: +81-6-6399-0008
Nagoya Office	7-430 Morioka-cho Obu Aichi Phone: +81-562-46-2105, FAX: +81-562-46-2106
Business Contents	1. Development support for drugs, diagnostic pharmaceuticals, and biomarkers, medical imaging techniques, and know-how 2. Clinical development support (clinical study monitoring, quality control, image analysis, image data handling, image data central review) and monitoring in clinical trials for medical drugs/devices 3. Operation support for PET tracer synthesis, PET animal imaging assessment, and PET manufacturing 4. Regulatory affairs consulting support (e.g. establishment of QA system, GMP/c-GMP for PET drugs) 5. Consulting services for drug development
URL	https://micron-kobe.com
Email	imagingbiomarker@micron-kobe.com

Micron Inc.

Reference

1. 「次世代医療機器評価指標の公表について　別紙4人工知能技術を利用した医用画像診断支援システムに関する評価指標」（令和元年5月23日付け薬生機審発0523第2号）

2. 「平成30年度次世代医療機器・再生医療等製品評価指標作成事業　人工知能分野　審査WG 報告書」（平成31年３月国立医薬品食品衛生研究所）

3. https://www.amed.go.jp/news/release_20181210.html

4. FDA. Clinical Performance Assessment:Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data -Premarket Approval (PMA) and Premarket Notification [510(k)] Submissions
https://www.fda.gov/media/77642/download

5. 「コンピュータ診断支援装置の性能評価開発ガイドライン2015（手引き）」(平成27年12月経済産業省／国立研究開発法人日本医療研究開発機構)

6. 杣取恵太・坂本次郎・時田椋子・鈴木彩夏・国里愛彦. 診断精度研究の系統的レビューとメタアナリシス. 専修人間科学論集　心理学篇　Vol. 6, No. 1, pp. 41～58, 2016

7. 「診断用放射性医薬品の臨床評価方法に関するガイドライン」（平成24年6月11日付け薬食審査発0611第1号）

8. Bertens LC, Broekhuizen BD, Naaktgeboren CA, Rutten FH, Hoes AW, van Mourik Y, Moons KG, Reitsma JB. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. PLoS Med. 2013 Oct;10(10):e1001531.

9. Klaus Gottlieb, Fez Hussain. Voting for Image Scoring and Assessment (VISA) - theory and application of a 2 + 1 reader algorithm to improve accuracy of imaging endpoints in clinical trials. BMC Med Imaging 2015 Feb 19;15:6.