眼底画像診断を補助するプログラム医療機器に求められる精度に関する調査報告書

2021 年 3 月　第 1 版

日本眼科 AI 学会

1，はじめに

第 4 次産業革命とも呼ばれる様に、近年の、ディープラーニングを中心とした人工知能の台頭は目覚ましい。2018 年 4 月には米国規制当局が、眼底画像から糖尿病網膜症の有無を判定する医師不要の AI プログラムを Software as a Medical Device (SaMD)として初めて承認している。ヘルスケア分野は第 4 次産業革命において最も注目されている分野の一つであり、この流れは加速している。

本邦においても、2018 年 12 月に機械学習（サポートベクターマシン）を活用したプログラム医療機器が初めて承認され、2019 年 10 月にはディープラーニングを用いたプログラム医療機器も承認されるなど、人工知能を活用したプログラム医療機器が日常診療で活用され始めている。眼科領域においても、光干渉断層血管造影のデノイジングにディープラーニングを活用した光干渉断層計（OCT）が既に販売されている。

このように人工知能を活用したプログラム医療機器が承認を得ているが、どの程度の性能であれば臨床的意義があると言えるのかについては常に課題となっている。プログラム医療機器を新たに作成する度に対照群も評価するという方式は医療機器の開発 にとってスピード・コストの観点から大きな足かせとなるほか、対照群が示す性能にもばらつきがあること、評価項目をセットする段階では臨床的な観点が必ずしも入らないことなどから適切とは言い難い。そこで、臨床的意義に基づき、臨床的な視点から一定程度妥当と考えられる水準を示すことが期待されている。

2，目的

本調査報告書は、本邦における眼底画像診断補助プログラム医療機器を開発するにあたって求められる精度を、日本眼科 AI 学会が実施した試験に基づいて考察し、一定の基準値となる値を提示するものである。

3，本報告書で取り上げるプログラム

(1) 健康診断施設等において眼底画像所見が正常であるか否かを診断する補助を行うプログラム

① 当該プログラムの用途

通常、健康診断施設においては、非散瞳下で眼底画像を撮影した後、当該データを眼科医（主に非専門医※）へ送信して読影を行うという方式を採用している。しかしながら、一般に正常眼が 90～95%程度を占め、明らかに正常と思われる眼底画像も眼科医が読影を行っていることから、健康診断施設および読影する眼科医双方の

負担となっている。このため、正常であるか否かの診断を補助するプログラムを用いて、健康診断施設の医師が明らかに正常な眼底画像を診断し、眼科医は異常所見を有する事前確率が高い眼底画像に注力する戦略が有効となる。これによって眼科医が読影を要する画像枚数を減らすことができて物理的な負担が軽減されると共に、異常所見を有する事前確率が高い眼底画像に注力できることから見落とし防止に寄与する。

※　眼科医（非専門医）とは、公益財団法人日本眼科学会が定める日本眼科学会専門医制度規則第 8 条（2）に定める経験を満たさない眼科医師を指す。

（日本眼科学会専門医制度規則第 8 条（2）　第 9 条に規定する施設において、施行細則で定める研修内容により 5 年以上眼科臨床を研修した者。あるいは厚生労働省の定める卒後臨床研修（2 年間）終了後、第 9 条に規定する施設において施行細則で定める研修内容により 4 年以上眼科臨床を研修した者。即ち卒後臨床研修を含め 6 年以上の臨床経験を終了した者）

② 日本眼科 AI 学会が行った試験の概要

日本眼科学会の構築する Japan Ocular Imaging Registry に収集された眼底画像につき、眼科医（非専門医）が眼底読影を行った。その結果を教師ラベルと比較してパフォーマンスを評価した。

③ 方法

2020 年 7 月 9 日から 7 月 16 日に、眼科医（非専門医）13 名が眼底画像の読影を行った。読影した眼底画像データセットは、日本眼科学会が収集し、国立情報学研究所に保管する 12 クラス 13,445 枚の眼底画像から構成比率を維持してランダムに抽出した 300 枚のデータセット。この 12 クラスは、以下の表 1 に示すように、眼科領域で頻度の高い疾患をカバーしている。各眼科医は各画像を 12 クラスに分類した。正常または異常（正常以外の 11 クラスいずれかへの分類）の 2 値判定における識別能力を、感度及び特異度で評価した。

表 1、眼底画像に付与されている 12 クラスの教師ラベル

正常	加齢黄斑変性（早期、後期）
中心性漿液性脈絡網膜症	網膜静脈分枝閉塞症 網膜中心静脈閉塞症
黄斑円孔	黄斑上膜
糖尿病網膜症（非増殖性、増殖性）	緑内障（前視野、初期、中期、後期）
近視性網脈絡膜萎縮 近視性新生血管	乳頭浮腫
網膜色素変性症	非緑内障性視神経萎縮

④ 日本眼科 AI 学会が行った試験の結果

異常を検出する感度及び特異度を表 2 にまとめる。13 名中 9 名（69.2%）は 90%以上の感度を認めたが、その 9 名の中でも特異度は 28.6%～77.1%とばらつきを認めた。一般に感度を高くすると特異度は低下するが、今回の試験結果でもその傾向が確認された。13 名の感度の平均値は 91.8%、中央値は 91.3%で、特異度の平均値は 54.1%、中央値が 51.4%であった。

表 2、13 名の眼科医（非専門医）の異常検出感度及び特異度

	感度	特異度
A	0.992	0.286
B	0.981	0.457
C	0.970	0.400
D	0.962	0.686
E	0.958	0.429
F	0.928	0.629
G	0.913	0.686
H	0.909	0.714
I	0.906	0.771
J	0.894	0.514
K	0.891	0.686
L	0.838	0.400
M	0.785	0.371
平均	0.918	0.541
中央値	0.913	0.514

⑤ 試験結果を踏まえた、求めるべき精度等についての考察

本プログラムの使用方法に鑑みると、異常所見がない（陰性）と判断された画像が正しく陰性である確率（陰性的中率）が十分に高いことが望まれる。健康診断における事前確率（有病率）を 5%または 10%とした場合の、異常検出感度別の陰性的中率（正常判定の眼底画像が実際に正常である割合）を表 3 に示す。

b

表 3、感度、特異度、有病率別の陰性的中率

感度	特異度					
	60%		70%		80%	
	有病率 10%	有病率 5%	有病率 10%	有病率 5%	有病率 10%	有病率 5%
99%	99.8%	99.9%	99.8%	99.9%	99.9%	99.9%
98%	99.6%	99.8%	99.7%	99.8%	99.7%	99.9%
97%	99.4%	99.7%	99.5%	99.8%	99.6%	99.8%
96%	99.3%	99.7%	99.4%	99.7%	99.4%	99.7%
95%	99.1%	99.6%	99.2%	99.6%	99.3%	99.7%
94%	98.9%	99.5%	99.1%	99.6%	99.2%	99.6%
93%	98.7%	99.4%	98.9%	99.5%	99.0%	99.5%
92%	98.5%	99.3%	98.7%	99.4%	98.9%	99.5%
91%	98.4%	99.2%	98.6%	99.3%	98.8%	99.4%
90%	98.2%	99.1%	98.4%	99.3%	98.6%	99.3%

眼科医（非専門医）の平均値である感度 91.8%、特異度 54.1%を用いた場合の陰性的中率は、有病率 5%で 99.2%（※有病率 10%では 98.3%）となることから、眼底画像が正常であるか否かを診断する補助を行うプログラムでは、有病率 10%を想定した場合であってもこの数値以上に陰性的中率が高くなることを示す必要があると考える。

具体例としては、(A) 特異度が 60%以上かつ感度が 95.4%以上であることを示す方法や、(B) 特異度が 70%以上かつ感度が 94.6%以上であることを示す方法、(C) 特異度が 80%以上かつ感度が 93.9%以上であることを示す方法などが考えられる。

なお、同程度の陰性的中率なのであれば陽性的中率も高い方が望ましいため、有病率 5%を想定した場合に陽性的中率が 11.2%となる(A)や同 14.2%となる(B)よりも、同 19.8%となる(C)の方がより優れていると判断される。

4, 用いるデータセットについて

用いるデータセットの規制科学的妥当性は医薬品医療機器総合機構（PMDA）が所管しているため、どのようなデータセットにより検証するかは PMDA との相談によって決定する。

日本眼科学会及び日本眼科 AI 学会の見解としては、今回調査した眼底画像に限らず、本邦で使用する AI 関連のプログラム医療機器については、日本人データを用いて検証されていることが望ましいと考える。また、検証データの信頼性を担保するため、治験や臨床性能評価試験等にあたっては、日本眼科学会、日本眼科 AI 学会ほか日本眼科学会関連学会と連携し、多施設のデータを用いることが望まれる。なお、その代替として、日本眼科学会が構築するデータベースである Japan Ocular Imaging Registry のデータを用いるのも一手である。いずれにしても、子細はまず日本眼科学会、日本眼科 AI 学会等に相談するのが望ましい。

5, 終わりに

本調査報告書では、眼底画像診断を補助するプログラム医療機器に求められる精度について報告した。AI プログラムは、判定する項目が同じであったとしても、その臨床的位置づけによって求められる性能が異なるため、本報告書に掲載されているプログラムと判定する項目が同じであったとしても、その用途によっては本報告書で示した参考値が必ずしも当てはまらないことには十分に注意が必要である。

今後も、求められる情報について継続的に調査を進め、本報告書を改訂する。

＜改訂履歴＞

2020 年 11 月　第 1 版