令和 5 年 8 月 28 日

独立行政法人医薬品医療機器総合機構
理 事 長　　藤 原 康 弘　殿

科 学 委 員 会
委員長 渡邉 裕司

科学委員会では、今般、下記について科学的見地からの議論をまとめました。
独立行政法人医薬品医療機器総合機構における通常業務にご活用ください。

記

AIを活用したプログラム医療機器に関する報告書

以 上

AI を活用したプログラム医療機器に関する報告書

目次

【用語一覧】

略語	英語名称	日本語名称
ACP	Algorithm Change Protocol	SPS で規定された変更の実施、検証及び妥当性確認方法を規定したアルゴリズム変更プロトコール
AI	Artificial Intelligence	人工知能
AMED	Japan Agency for Medical Research and Development	国立研究開発法人日本医療研究開発機構
AUC	Area Under the Curve	曲線下面積
CNN	Convolution Neural Network	畳み込みニューラルネットワーク
DL	Deep Learning	深層学習
DICOM	Digital Imaging and Communications in Medicine	医療におけるデジタル画像と通信
FDA	U.S. Food and Drug Administration	米国医薬品食品局
GMLP	Good Machine Learning Practice	機械学習の安全性、有効性及び品質を確保することを目的とした管理
IDATEN	Improvement Design within Approval for Timely Evaluation Notice	医療機器変更計画確認申請制度
IEC	International Electrotechnical Commission	国際電気標準会議
IID	independent and identically distributed	独立同一分布
IMDRF	International Medical Device Regulations Forum	国際医療機器規制当局フォーラム
ISO	International Organization for Standardization	国際標準化機構
ML	Machine Learning	機械学習
MLMD	Machine Learning-enabled Medical Device	機械学習を利用した医療機器
PEMS	Programmable Electric Medical Systems	プログラマブル電気医用システム
PMDA	Pharmaceuticals and Medical Devices Agency	医薬品医療機器総合機構
SaMD	Software as a Medical Device	医療機器プログラム

SPS	SaMD Pre-Specifications	機械学習を利用した SaMD を管理するために、その性能、入力、使用目的に対する計画的な変更を規定した事前仕様
WSI	Whole-Slide Imaging	病理デジタル画像

１．はじめに

人工知能（Artificial Intelligence; AI）を活用した医療機器ソフトウェアに関する科学委員会の報告に関しては、2017 年に「AI を活用した医療診断システム・医療機器等に関する課題と提言 2017」[1]を公表し、2017 年時点での AI 技術の現状を概観し、その技術的側面を解説するとともに、AI 医療システム[1]のレギュラトリーサイエンスに関する基本的な論点及び AI 医療システムの倫理・責任に関する議論等についても示してきた。

その後、多くの AI を利用した医療機器が承認され、2021 年には、医療機器プログラム（Software as a Medical Device; SaMD）の開発に係る事業の予見可能性を高めることを目的として、「プログラムの医療機器該当性に関するガイドライン」[2]が発出、2023 年 3 月に一部改正された[2]。機械学習（Machine Learning; ML）を応用した医療機器の特徴である、市販後に実臨床でのデータを学習することにより性能が変化するという特徴を有する医療機器の実用化への期待は大きいが、その開発は未だ限定的であると推察する。

一方、各国の規制当局においても市販後性能が変化する性質を有する医療機器に関する議論が進展しており、ステークホルダー間の議論が活発となっている。

本報告で対象とする AI を活用した医療機器としては、医用画像解析、心電図、脈波などの各種生体計測情報解析分野、放射線治療計画立案支援などを想定した。なお、ゲノムデータ解析分野については対象としない。また、本報告作成時点で、自然な対話形式で AI が答えるチャットサービスが登場し、医療従事者が自然言語で発する質問に対して回答する AI 医療システムの研究開発も実際に行われているが、本技術についてはいまだその特性が未解明であること、誤った回答を AI が発することなどが指摘されており、現段階で議論することは時期尚早と判断し本報告の対象からは除外した。

本報告では 2017 年の報告以降、現実の課題となってきている以下の課題に着目し、専門部会として議論を進め報告書としてとりまとめることとした。

- 国内外の医療機器規制・医療機器安全規格制定活動の動向分析
- 機械学習におけるバイアスの諸問題
- 市販後学習における評価データの再利用に関する問題と問題解決に向けた研究の現状
- 物理モデル・シミュレーションによる学習データ構築の現状と課題
- 学習データ、検証データ、テストデータのデータソースとしての各種臨床情報データベースに関する諸課題

---

[1] 「AI を活用した医療診断システム・医療機器等に関する課題と提言 2017」においては、AI を構成要素として含む疾患の診断、治療又は予防に使用されることが目的とされているシステム（薬機法上の医療機器に限定されてない）を「AI 医療システム」と称している。

[2] 薬機法に基づき規制される医療機器プログラムは、医療機器としての目的性を有し、かつ意図したとおりに機能しない場合に患者（又は使用者）の生命及び健康に影響を与えるおそれがあるプログラム（ソフトウエア機能）とされている。

国内外の医療機器規制・医療機器安全規格制定活動の動向分析では、Good Machine Learning Practice（GMLP）の概念や、性能変化に関する規制の視点からの論点を議論した。

機械学習におけるバイアスの諸問題では、一般に学習（Training）、検証（Validation）、テスト（Test）の 3 つのデータが現実に応用を想定する患者群の統計的性質と類似しているかという観点から、「バイアス」の問題が語られることが多い。また、このようなデータのバイアスのみならず、使用状況に関するバイアスなど、機械学習を応用した AI 医療機器の開発を進める上で注意すべき「バイアス」について議論した。さらに、医用画像解析への深層学習（Deep learning; DL）に関して、現実に存在するこのようなバイアスの例について文献を引用して紹介した。

市販後学習における評価データの再利用に関する問題及び問題解決に向けた研究の現状では、理論的には学習データ・検証データとテストデータを独立に収集し、そのコンタミネーションを防ぐ措置が取られていなければならない。一方、評価結果を知った上で、この弱点を解消する目的で意図的に又は意図せずに学習データを偏って収集して学習を行うなどの不適切な開発手法がとられた場合には、テストデータが独立に収集されていたとしても過学習の可能性があることなどを指摘し、関連する論点を整理した。

さらに、学習データ生成の手法として、数値シミュレーションを活用する手法の現状及び課題について議論を行い、2021 年に公表した科学委員会報告書「コンピューターシミュレーションを活用した医療機器ソフトウェアの審査の考え方に関する専門部会報告書」[3]の議論等を参考に関連する課題の論点を整理した。

学習データ、検証データ、テストデータのデータソースとしての各種臨床情報データベースに関する諸課題では、具体的に手術画像データ、内視鏡データ、病理画像データ、心電図データのデータベースを用いて機械学習を応用した AI 医療機器の開発を行う際に、データベースが有するべき共通の性質と、分野固有の課題に関する論点を整理した。

なお、実際の臨床において学習データ、検証データ、テストデータ等を収集する際には、個人情報保護の課題が現実の研究開発で大きな課題となるが、これは法的な課題で科学委員会による科学的な見地に基づいた議論には馴染まない問題であり、本報告のスコープを超えている。しかし、現実の研究開発においては重要な課題であることから、現実のデータベース開発上の課題を簡単に記述した。またこれらの貴重なデータを医療機器ソフトウェア開発に活用する上で、最低限考慮しなければならない倫理的な観点からの要求について簡単に言及した。

２．国内外の関連動向の分析

２－１．国内外の関連動向

米国（U.S. Food and Drug Administration; FDA）

FDA は、「医療機器プログラムに基づく人工知能／機械学習（AI/ML）の変更に対応する規制の枠組み」に関するディスカッションペーパーを 2019 年 4 月に発行した[4]。その目的の一つは、市販後に性能を変化させることが可能な AI、特に機械学習を利用した SaMD の開発及び実用化促進であり、そのために必要となる要件を整理することだったと推測される。事実、このペーパーには、これまで性能を「固定」した場合のみ承認していた機械学習を利用した SaMD において、その特徴である追加学習に伴う性能変化を実装した製品を承認可能とするためには新しい枠組みが必要であると言及されている。そのため、このペーパーでは安全性及び有効性の維持に資する品質マネジメントシステムと市販後調査に重点をおいた Total Product Life Cycle アプローチを新しい枠組みに適用することが提案されている。そのために必要な枠組みとして品質システムのみならず、GMLP の導入等も含めた一般原則を記している。

上述のペーパーで、FDA は追加学習に伴う性能変化が可能な機械学習を利用した SaMD を管理するために、その性能、入力、使用目的に対する計画的な変更を規定した事前仕様（SaMD Pre-Specifications; SPS）、並びに SPS で規定された変更の実施、検証及び妥当性確認方法を規定したアルゴリズム変更プロトコール（Algorithm Change Protocol; ACP）の両者を含んだ変更管理計画を作成・使用することを推奨している。その後、その推奨事項をより実装可能な形に発展させ、2023 年 4 月、FDA は AI/ML を利用した SaMD（AI/ML Enabled Medical Devices）の機能に関する事前変更管理計画を含んだ承認申請に関する推奨事項に関するガイダンス案を発行した[5]。

また、FDA は、寄せられたコメントや具体的な対応策に関する検討等の取り組みにより、製造業者による透明性と real world における性能のモニタリングを達成することを期待していた。その後、2021 年 5 月に寄せられたパブリックコメントと対応策案を取りまとめたアクションプランが発行され、今後の取り組みにおける 5 つの論点が示された[6]。

1. AI/ML を利用した SaMD に対する規制の適切な枠組み（変更管理計画に関するガイダンスの必要性を含む）
2. GMLP
3. 使用者への透明性を確保する患者中心のアプローチ
4. アルゴリズムのバイアス及び堅固性に関するレギュラトリーサイエンス研究
5. リアルワールド性能と実際に得られたデータ（real world data）の活用

その後、これらの論点に対する回答を検討する目的で、FDA が把握している AI/ML を利用した SaMD リストの公表、AI/ML を利用した SaMD の透明性に関するワークショップの実施、GMLP 指導原則の発出（カナダ及び英国と共同）等が実施されてきている。

なお、FDA のウェブサイトで公開された AI/ML を利用した SaMD の承認リストは、2022 年 10 月 5 日、1 年以上ぶりに更新された[7]。更新前との比較から、新たに 155 製品が承認されたことが明らかとなったが、本邦における「一部変更」に相当する承認品、単なるイメージデータの改善や他のシステムに使用するためのフォーマット変更を目的とした製品、510(k) summary においても使用アルゴリズムを明記していない製品もリストに存在していることが確認された。また、機械学習を利用した SaMD と古典的アルゴリズムを利用した従来型の SaMD とを製品コード等で区別しておらず、主機能に機械学習を利用した SaMD の承認数を正確に把握することは困難であった。

FDA は 2017 年に SaMD の審査促進を目的とした Pre-certification プログラム（Pre-Cert）の試験的導入を発表し、2019 年 1 月に関連する文書 3 つを発表して試験を開始した。2022 年 9 月にその成果物を公表したが、Executive summary において、Pre-Cert の試験的運用が完了した一方で、その過程で直面した種々の課題を考慮した更なる規制枠組みの開発検討を継続することが記載されている[8]。課題として挙げられた事項は以下のとおりである。

- Pre-Cert のような組織的評価により市販前審査を不要にできる低リスクのデバイスを特定するには未だ手法開発が不十分である
- 中適度のリスクを有するデバイスにおける固有の臨床性能評価及びサイバーセキュリティ評価を置き換えるには不十分である
- デバイス固有の評価が特に高程度のリスクを有するデバイスにおいては極めて重要である

すなわち、Pre-Cert による審査促進については、低リスクのデバイスを特定可能な要件の整理とその判断方法の明確化、中高程度のリスクを有するデバイスの特徴を把握した上で、個別の留意点を整理した評価方法の確立が必要であるが、現時点の科学的知見に基づいた上記課題に対する解決策の構築が困難であり、更なる検討が必要であることを示している。

これらの施策に加え、FDA は、現在、SaMD を含むデジタルヘルス関係の医療機器に関する課題の取りまとめや将来の活動を担当する Digital Health Center of Excellence を医療機器・放射線保険センター（Center for Devices and Radiological Health; CDRH）直下に設立しているが、承認審査への関与程度について判断する情報は把握できていない。

注 : 国立医薬品食品衛生研究所医療機器部の調査研究報告書に基づいて取りまとめた内容であるため、詳細はそちらを確認のこと[9]。

米国、カナダ、英国による共同文書

米国、カナダ、英国の規制当局である FDA、ヘルスカナダ、MHRA（Medicines and Healthcare products Regulatory Agency）は、AI/ML を利用した SaMD の安全性、有効性及び品質の確保を達成することを目的に GMLP として 10 の指針を示す共同文書を 2021

年 10 月に発行した[10]。これらの指針は、国際医療機器規制当局フォーラム（International Medical Device Regulations Forum; IMDRF）、規格作成団体、学会等が GMLP を基本として議論を展開することを期待したものである。示された具体的な指針は以下のとおりである。

1. Total Product Life Cycle を通じて、多分野の専門家が参画する
2. 適切なソフトウェア工学及びセキュリティ対策を適用・実施する
3. 臨床試験の参加者及びデータセットが意図する患者集団を代表する
4. 学習用データセットがテスト用データセットと独立している（独立性確保）
5. 最も適切な手段に基づき入手、選別された正解ラベル付きデータセットを確保する
6. 入手可能なデータに適合し、医療機器の使用目的を反映したモデルを設計する
7. 人間と AI がチームとなった状態での性能に重点を置く
8. 試験では臨床的位置づけを加味した医療機器の性能を実証する
9. 使用者に明確で、基本的な情報を提供する（透明性の確保）
10. 上市されたモデルの性能をモニタリングし、再学習のリスクを管理する

欧州連合（EU）

2021 年 4 月、欧州委員会は EU 市場における AI を利用した製品・サービスに対する規制枠組み案を発表した。正式名称は「人工知能に関する整合的規則（人工知能法）の制定及び関連法令の改正に関する欧州議会及び理事会による規則案[11]」である。この規則案は欧州医療機器規則（Medical Device Regulation; MDR）と欧州体外診断用医療機器規制（In Vitro Diagnostic Regulation; IVDR）により規制される医療機器等も対象となっている。規則案の目的は

- 欧州連合市場に上市される AI システムが安全であり、基本的権利に関する既存の法律及び欧州連合の価値観を尊重・保証すること
- AI への投資とイノベーションを促進するために法的確実性を確保すること
- AI システムに適用される基本的権利と安全要件に関する既存の法律による管理と効果的執行を強化すること
- 合法的で安全かつ信頼できる AI アプリケーションの単一市場発展を促進し、市場の断片化を防ぐこと

とされている。当該規則案は最新状況を踏まえた議論を経て、2023 年 5 月、欧州議会の関連委員会で承認された。2024 年以降の施行を目指し、欧州議会等での議論が進んでいる。

この規則案では、リスクベースアプローチが採用されており、AI を利用した SaMD は高リスクシステムとして規制される方向であるが、COCIR 等の医療機器団体は規則案をそのまま法制化することに反対の立場をとっている。欧州で設立された AI 高度専門家グループも、一般的な政策や規制を幅広く適用することは特定分野においては有害

となり得ることを公表している。事実、規則案が MDR、IVDR と整合しない要件を含んでいることが医療機器業界より指摘されている。

注：詳細は国立医薬品食品衛生研究所医療機器部の調査研究報告書を参照のこと。

韓国

政府主導で 2015 年より AI に関連する施策が開始されており、2019 年に「人工知能国家戦略」が公表されるまでの間、複数の戦略が公表されている。また、関連するガイドラインも多数公表されている。

「人工知能国家戦略」は AI 時代に向けた国家的ビジョンと政策を関係省庁が連携して作成した国策であり、2030 年までに関連する競争力において世界 3 位の地位を確保すること、AI により 455 兆ウォンの経済利益を創出すること、結果として世界上位 10 ヶ国に入る生活の質を確保することを目標として挙げている。

2019 年の IMDRF ロシア年次総会では、AI 利用製品の大幅な増加を見込み、その迅速な市場参入のための規制緩和を実施していることを公表した。

注：詳細は国立医薬品食品衛生研究所医療機器部の調査研究報告書を参照のこと。

中国

2019 年の IMDRF ロシア年次総会において、国内ガイダンス案「深層学習による意思決定支援 SaMD の審査ポイント」を紹介した。その際、FDA と同様、当該製品のトータルライフサイクルマネジメントに重点を置いた取り扱いを考慮していることが強調されている。このガイダンス案は 2022 年 3 月に公表予定となっていたが、現状は不明である。

2022 年 5 月の段階で、9 つの関連ガイダンスが発行されていることが確認できている。

注：詳細は国立医薬品食品衛生研究所医療機器部の調査研究報告書を参照のこと。

日本

医療機器規制に資する活動としては、厚生労働省の次世代医療機器・再生医療等製品評価指標作成事業における、AI 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案作成を目的とした審査 WG の設立が挙げられる。2017 年度から 2018 年度まで活動した審査 WG の成果物は、パブリックコメントを経て 2019 年 5 月に厚生労働省医薬・生活衛生局医療機器審査管理室長通知として発出され評価指標となった[12]。

また、2022 年（令和 4 年）6 月 7 日に閣議決定された規制改革実施計画に基づき、類型ごと、対象疾患ごとに承認実績がある SaMD を認証制度に移行するために、認証基準

の策定及び改正が進んでいる。並行して、独立行政法人医薬品医療機器総合機構（Pharmaceuticals and Medical Devices Agency; PMDA）においては、開発事業者の予見性を高めるために有効性・安全性評価に必要な試験条件や評価ポイント等、審査ポイントに関する情報の整理とウェブサイトを介した公表を開始している[13]。

科学的側面からの支援活動としては、2019 年より国立研究開発法人日本医療研究開発機構（Japan Agency for Medical Research and Development; AMED）医薬品等規制調和・評価研究事業「人工知能等の先端技術を利用した医療機器プログラムの薬事規制のあり方に関する研究」が挙げられる。当該研究では、産学官連携の下、市販後の追加学習が可能な AI を利用した SaMD の実現可能性が検討された結果、製造販売業者による段階的な追加学習と性能変化を既存の規制枠組み、特に医療機器変更計画確認申請制度（Improvement Design within Approval for Timely Evaluation Notice; IDATEN）を利用して実施するための提言が作成され厚生労働省に提出された。また、市販後の追加学習を経た SaMD の性能に影響を与える学習データの要因を明らかにするための実証研究も行われ、その成果も提言に組み込まれた。

現在、上述した研究の後継事業として、2022 年より AMED 医薬品等規制調和・評価研究事業「AI を利用した医療機器プログラムの市販後学習時の性能評価に資する研究」が進行しており、性能評価プロセスの妥当性を判断する上で必要となる留意点を検討するための実証研究が進んでいる。今後、産学官連携体制を構築し、実証研究の成果も踏まえた性能評価ガイダンス案の作成を実施する予定となっている。

一方、厚生労働科学研究費補助金事業による検討の結果が取りまとめられ、SaMD の特性を踏まえた承認及び開発ガイダンスとして厚生労働省より 2023 年 5 月に発出された[14]。今後、規制当局による施策の成果として、SaMD の社会実装に資する文書が公となっていくことが期待される。

IMDRF

IMDRF において発足した人工知能を取り扱う AI WG は、機械学習を応用した医療機器（Machine Learning-enabled Medical Device; MLMD）の用語に関するガイダンス文書を 2022 年 5 月 6 日に発行した[15]。今後の方針についても WG 内で議論が継続されており、将来的な取り組みとして以下の事項を検討し、ガイダンスを発行することを予定している。

1. GMLP（2023 年～2024 年）
2. 事前変更計画（Predetermined Change Control Plan; PCCP）（2024 年～2025 年）
3. 上記議論で生じた課題（2025 年～）

なお、別途データマネジメントについても将来的に検討することも予定している。

IEC/TC 62/SNAIG

IEC/TC 62/Software Network and Artificial Intelligence advisory Group（SNAIG）は、AI 又

はそれに関連する事項を取りまとめ、今後の医療機器に関する規格策定の指針を示すアドバイサリーグループである。

SNAIG は、AI 利用医療機器分野における規格を以下の 3 層構造とすることを提案している。

- 基礎規格（AI Base）

  医療機器として要求される ISO 14971（リスクマネジメント）、ISO 13485（品質管理マネジメント）、IEC 62304（医療機器プログラムのライフサイクル）、IEC 62366-1（医療機器のユーザビリティエンジニアリング）等も考慮した、AI の基礎となる要求事項を記載した規格が想定される。

- 機能規格（AI functional）

  例えば、画像解析や波形解析等、特定の機能と関連する規格が想定される

- 意図した用途に応じた手法、あるいは評価に特化した規格（AI particular）

  例えば、SaMD であれば IEC 82304-1（医療機器プログラム）に加えその用途に特化した規格、プログラマブル電気医用システム（programmable electric medical systems; PEMS）であれば IEC 60601-1（医用電気機器）に加えその用途に特化した規格などが想定される。

AI base 及び AI functional の国際規格はまだ存在しないが、国単位では同様の要求事項を含めた規格が作成、あるいは提案されていることを考慮して、SNAIG は、IEC/TC 62 で開発する人工知能を活用した医療機器の規格には、以下 3 つの観点を付属書に追加することを推奨している。

1. 基礎となる要求事項（AI base 関連事項：バイアスの管理、試験方法等）
2. 機能に関する要求事項（AI functional 関連事項：画像解析、波形解析等）
3. 個別の要求事項（AI particular 関連事項：歯科領域の X 線画像解析、舌画像解析等）

今後、IEC/TC 62 において、独立した学習データとテストデータの枚数を、標榜する内容を達成するために、どのような統計的方法を用いて提供可能かを検討するための予備評価項目（Preliminary Work Plan; PWI）が開始される予定である。

一方、医療情報を扱う ISO/TC 215 にデータマネジメントに関する規格の策定、医療機器全般を取り扱う ISO/TC210 に人工知能を活用した医療機器の特性を反映する ISO 13485 を改正することを SNAIG は要望している。

なお、前述した欧州における人工知能法の 2024 年中の施行を考慮した結果と思われるが、その施行に利用するための ISO（国際標準化機構）及び IEC（国際電気標準会議）における AI 関連規格の作成が進み始めている。この報告書が公開される頃には、医療機器関連の規格が次々と提案、作成されていることも十分考えられる。

２－２．国内外動向に関する分析と考察

各国の規制状況は、AI に対して性善説のスタンスを取るか、あるいは性悪説のスタ

ンスを取るかで異なっている。AI 技術の発展を促進したい国は前者のスタンスとなっていることが見受けられるが、EU は後者のスタンスを取り、SaMD のみならず各種技術への AI 適用を規制しようとしている。過度な規制は技術発展を阻害しその利便性供与も受けなくなる可能性があることを EU は理解していると思われるが、それ以上に個人情報の漏洩や過度なイノベーションによる個人の権利へのリスク等を強く懸念していることが窺える。その背景には、現在開発されている AI において、その内部アルゴリズムが明確でなく、責任の所在が曖昧になりかねないという点があると思われる。上述した背景は、性善説を取る他の国においても共通であり、その解決に向けたスタンスが異なるだけだと考えられる。よって、科学的な課題のみならず、社会的課題の解決も考慮することが肝要である。しかしながら、本専門部会は、機械学習を応用した SaMD に対して、科学的視点に基づいた適切な薬事規制枠組みを検討するための議論を進めるために設置されたことから、社会的課題の解決についてはその必要性を喚起するのみに留めることとする。

各国の動きを見ると、古典的アルゴリズムを使用したプログラムと、現在大きな着目を浴びている機械学習（深層学習を含む）を使用したプログラムとが分類されないまま、AI を包括的な用語として使用されているケースが散見される。一般的な用語として使用する分に問題は生じないのかもしれないが、医療機器規制において解決すべき様々な課題が生じていることを鑑みると、機械学習を応用したプログラムであることを改めて明確にする必要がある。

例えば、AI 技術において先端を走っていると思われる米国においては、機械学習の利用を前提とした「AI/ML」という表記が AI の代わりに使用されつつあるが、機械学習利用の有無が判別可能な製品コードが存在しないなど、AI の区別を逡巡している印象がある。一方、IMDRF は発行した文書[15]で明確に「MLMD」という用語を定義・使用しているが、その用語が積極的に世界各国で使用されている様子がない。本邦においても、MLMD という用語は使用されておらず、追加学習によりその性能が変化し得る機械学習ベースの AI 医療機器と古典的アルゴリズムによるプログラムとの区別が曖昧になっていることが多い。場合によっては、その区別が曖昧なままそれらに必要な薬事規制枠組みが議論対象として取り上げられることがあり、結果としてその議論が収束しないことがある。科学技術の国際整合の場である ISO 及び IEC で用語統一が進み各国でその用語が導入されることを期待したいが、現在の討議は機械学習ベースであることを前提に AI を定義しないまま議論が進んでいるような印象を受ける。内部では討議対象に関するコンセンサスが得られていたとしても、完成した国際標準が想定とは異なる形で利用される可能性があるため、その内容を外部に対して提示しておくこと、すなわち欧米が指摘している「透明性」を考慮することが重要である。そう考えると、FDA がアクションプランで提示した 5 つの論点、EU における規制案の目的は、国内において枠組みを考慮する上で参照すべき事項となり得る。それらの論点を踏まえ、AI 医療機器の薬事規制を構築するためには、まず、AI という用語を整理し、機械学習を利用して

対象疾病画像における特徴を検出する診断支援プログラムを代表とする機械学習を応用した SaMD が対象となることを明確にした上で議論を進めることが必要である。なお、機械学習を利用した SaMD の審査に必要となる各種評価においては、学習や性能評価で使用したテストデータの適切性や、その選定におけるバイアスの有無、使用する機器が異なる場合の影響等、患者に対するリスク要素を考慮する必要があるが、それらの詳細については後述の別項を参考にされたい。

機械学習を利用した SaMD の最大の特徴として、製品として上市した後も新たなデータを追加学習させることで性能を向上させることが可能なことが挙げられる。その特徴を最大限に活かした製品を実現化するため、FDA は Pre-Cert を試行したと思われる。この動きも AI に対する性善説が前提となっていることが想像され、非常に興味深い試みである。2023 年に公表された報告書[14]にはこれまでに明らかとなった課題が抽出されていると同時に、その試みを継続検討する必要性が述べられていることから、今後の展開が期待される。一方、前述したように、現時点での科学的知見では、開発され得る機械学習を利用した SaMD の市販後性能変化や適用拡大の範囲等を想定することが困難なだけでなく、実際に性能を変化させた場合のリスクに応じた製品の分類が難しく、その判断基準や要件を一様に設定することが現実的でないことから、結果的に、市販後の性能変化によるリスク・アンド・ベネフィット評価をメーカーに委ねることは難しく、その都度、規制当局が審査することが現時点では現実的だと FDA は判断したことも窺える。本邦においてもその特徴を活かした製品の実現を目指すメーカーが存在すると考えられるが、現時点では、米国同様に性能変化毎の審査が現実的な対応策になると考えられる。その際、IDATEN の積極的な活用を進めることが本邦における ML 利用 SaMD の社会実装迅速化に繋がることも期待される。

３．機械学習におけるバイアス

統計学においては、バイアスとは「標本が母集団の持つ様々な特徴からどの程度離れているか？」を示す用語として用いられる。一般的には母集団を得ることは難しく、標本から母集団を推定する必要がある。機械学習においては、一般的に有限の標本を用いて学習やその評価などが行われる。学習や評価に用いるデータが、対象とする問題の母集団からどの程度外れているのか、あるいは、学習、評価、そして、実際の臨床の場で利用されるデータとの間でそれぞれどの程度の差があるかの「バイアス」に関して正しく認識した上で研究開発を進める必要がある。また、バイアスがあるデータによって学習された機械学習モデルを用いて推論などを行うと、よりバイアスが強調される「バイアス増幅」なども知られている[16]。

データに関する統計学的な観点からの利用するデータに関する「バイアス」以外にも、いくつかのバイアスが考えられる。基本的には、①データ自身が持つバイアス（ある目的で利用するデータが母集団を正確に表しているかどうか)、②解析手法が持つバイアス、③認知バイアス（人が持つもの)、の 3 つがあると考える。以下、これらの観点からバイアスについて議論する。

３－１．データ自身が持つバイアス

画像などのパターン情報を機械学習などによって分類する場合、基本的には利用されるデータは、母集団を厳密に表したものである必要がある。機械学習による方法では、学習データから母集団が持つ特徴の分布を陰か陽に推定する。学習過程において利用しなかったデータを分類する場合には、この未知データが機械学習によって推定された特徴分布においてどのあたりに位置するかによって分類結果が出力される。したがって、分類対象となるデータは学習データが持つ特徴分布と一致している必要がある。しかしながら一般には、母集団を厳密に表したものを学習などにおいて用いることは不可能であり、何通りかのサンプリングによって得られる標本を用いることになる。このサンプリングでは、

1. 特定の病院のみで得られたデータ（特定病院でサンプリングされたデータ）
2. 特定の機器で撮影されたデータ（特定撮影機器でサンプリングされたデータ）
3. 特定の病態・年齢・人種などから撮影されたデータ（特定集団からサンプリングされたデータ）
4. 手法開発におけるデータ分割によって得られるデータ（母集団の特徴を正確に表さないデータ分割（一種のサンプリング）によって得られるデータ）

などが含まれよう。

1 は、特定の病院で得られたデータを基に学習を行った AI 医療機器が開発された場合には、他の病院において十分な性能を発揮しないことがあり得る。これは、学習データを生成した病院と AI 医療機器を利用する病院との間で、データ間にバイアスがある

ためである。大学病院で得られるデータと市中のクリニックで得られるデータの分布の間にも差がある。

2 は、特定の機器で撮影されたデータ（例 : A 社 CT 装置で撮影された胸部 CT 画像）を用いて学習を行った AI 医療機器を、異なる機器で撮影されたデータ（例 : B 社 CT 装置で撮影された胸部 CT 画像）では十分な性能が達成されない場合がある。これも、学習データを生成した機器で生成される画像と AI 医療機器を利用する病院で利用する機器から生成される画像との間で、データ間にバイアスがあるためである。画像撮影機器間の差は人の感覚的にはわからないことも多く、十分な注意が必要である。

3 は、特定の病態・年齢・人種・性別などから撮影されたデータを用いて AI 医療機器を構築した場合にも、それ以外の集団から得られるデータに対して十分な性能が達成できない場合がある。疾患の特徴は、国単位で異なることが多い。例えば、A 国で構築した AI 医療機器を B 国で展開した場合に、想定した性能を発揮しないこともある。

4 は、手法開発におけるデータ分割などによってもデータバイアスが発生する場合がある。N 分割交差検定などを用いて AI 医療機器開発を行う場合でも、分割後のデータ間で偏りが生じないように十分な注意を払う必要がある。

母集団を標本化（サンプリング）することによって得られるデータにはバイアスがあることを十分に認知した上で、臨床の場で分類すべきデータ（実世界データ）と学習データ（実験室データ）との間での偏りをできる限り少なくするような取り組みを実施することが必要であろう。これには使用する場面・条件などを限定する、使用環境を限定することなどで、バイアスによる影響を低減することも可能である。

３－２．解析手法が持つバイアス

画像認識手法自体にもバイアスがある。あるタスクに関して、「〇〇しやすい傾向がある」、「△△となる傾向がある」というものである。学習型の手法の場合には、データ自身が持つバイアスによって生じる場合もある。畳み込みニューラルネットワーク（Convolution Neural Network; CNN）の構造として様々な形のものが提案されているが、その構造によって抽出可能な構造物の大きさに偏りが生じる場合がある。これは、同じような学習データを用いても、CNN 構造によって抽出できる腫瘍の大きさに差が生じる。また、画像処理手順によって構築された手法であれば、その手順が持つ傾向によって結果に差が生じる。利用手法が持つバイアスをできる限り軽減するには、AI 医療機器の開発過程において、様々な手法、それが持つハイパーパラメータを試すことが必要であろう。

３－３．認知バイアス

気が付かないうちにデータの選択、手法の選択を行ってしまう人為的なバイアスも考えられよう。どのような病院からデータを選択するか、大量のデータから学習に必要なデータをどのように選択するかにおける選択バイアスが無意識に発生することが考え

られる。データ選択する行為自体、その行為を行う開発者が持つ専門知識によって左右される。例えば、大学病院だけのデータを用いるように開発計画を立てる、海外展開を考えているにも関わらず特定の国のデータのみを用いた開発計画を立てる、などがあろう。また、機械学習を用いた画像分類等の教師データ作成に関するバイアスもあろう。医師が持つ認知バイアスにより教師データにバイアスが生まれ、それを用いて機械学習による分類器を学習させることで、最終的に AI 医療機器がバイアスを持つ可能性もある。例えば、文献[17]では、新型コロナウイルス感染症（COVID-19）のコンピューター画像診断（Computed Tomography; CT）における病変領域の同定における医師が持つ認知バイアスについて示している。あるいは、画像から得られる所見のみで付与されたラベルから生じるバイアスもあろう。COVID-19 CT 画像のためのラベル生成を画像所見のみで行うのか、それとも PCR 検査に結果に基づくかの違いによって生じるバイアスもあろう[18]。これらは、無意識によって生じるバイアスであり、AI 医療機器開発には、十分な注意が必要である。そのために、AI 医療機器開発に当たっては、バイアスに関する総合的な教育を受けることも必要であろう。

４．市販後学習における評価データの再利用に関する問題と問題解決に向けた研究の現状

４－１．市販後再学習の性能評価の必要性

最近の AI、特に機械学習を利用した SaMD は、連続学習[19]などの頻繁な再学習が可能である。この特徴を利用することで、異なる施設間でデータの特徴が変化する Domain shift[20]等により市販後の SaMD が承認時の性能を発揮できない場合にも、市販後施設の学習データを用いて再学習することで性能向上が期待される。わが国では、このような SaMD の特徴を活かすために、変更計画も含めた承認が可能な IDATEN が整備されたが[21]、その活用はまだ十分ではない。その理由の一つに、市販後再学習により、性能は向上するだけでなく、低下する可能性もあること[22]、Catastrophic forgetting（破滅的忘却）[23]などのリスクが懸念されることが挙げられる。さらに、再学習を繰り返し行った場合の性能の評価の際には、同一テストデータを繰り返し利用する場合に潜むリスクに注意する必要がある。

性能の変化、特に性能の低下には適切に対処しなければならないため、再学習後の性能評価は重要である。この性能評価は、①市販前の承認時のテストデータを用いて行う場合、②市販後の新しいテストデータを用いて行う場合、の 2 通りが考えられる。市販前の承認時のテストデータを用いた評価は、破滅的忘却などの問題が生じず、承認時の性能が担保されていることを確認するために重要である。市販後のテストデータを用いた評価は、実際にシステムを運用する際の性能の確認のために必要となる。

４－２．市販後再学習の性能の評価法の例

評価の方法としては、市販前の性能に対する市販後の性能の非劣性又は同等性の評価[24]が考えられる。評価にあたっては、評価項目と許容範囲（マージン）を設定する必要がある。評価項目は SaMD の目的によって異なるが、例えば、画像の分類であれば Receiver operating characteristic (ROC) curve の下面積である Area Under the Curve (AUC) や Accuracy、領域の検出であれば正解領域と検出領域の間の Dice score や Intersection over Union (IoU) などの一致度、回帰であれば回帰誤差、などが想定される。マージンは臨床的見地などから適切な値を設定する必要がある。これらの評価項目とマージン値を決定した後は、例えば非劣性試験の場合には、テストデータを利用して信頼区間を推定し、マージンとの関係に基づいて検証を行う。

４－３．市販後再学習の性能評価においてリスクのある再学習法の例とその対処法

市販前の SaMD の性能は、通常、統計的に十分な数のテストデータにより評価される。市販後の再学習による評価においても十分な数のテストデータを用いて評価を行うことが望ましい。また、再学習を繰り返す場合には、毎回、再学習にまったく関与していないフレッシュかつ十分な数のテストデータを用いて性能を評価することが理想的

である。しかし、評価のための正解データ（アノテーション）の作成には専門知識や別の検査が必要になることもあり、時間的・経済的コストが高く、作成は容易ではない。そのため、限られたテストデータを再利用せざるを得ない状況が生まれる可能性が高い。

この同一テストデータを再利用した評価には、再学習の方法によっては、フレッシュなテストデータによる評価値（真の評価値）と異なってしまうリスクが存在する。例えば、リスクがある再学習法の例としては、同一テストデータによる評価結果を機械学習の分類器の設計に利用する場合や、複数の分類器を同一テストデータに適用してもっとも性能が高いものを選ぶなど、モデル選択のプロセスに同一テストデータによる評価結果を用いた場合、などである。いずれも、テストデータが実質的に学習プロセスの一部に組み込まれてしまい、フレッシュなテストデータで評価した場合とは異なり、評価値にバイアスが含まれてしまう。このバイアスを含んだ評価値を用いて非劣性試験を行っても、正しい結果は得られない。以下では、Dwork らが示した問題例や解決策[25]を通じて、テストデータへの過剰適合により真の評価値からバイアスが混入するリスクがある再学習法の例とその対処法について解説する。なお、特定のデータセットに対して過剰適合すると、そのデータセットによる評価値と、フレッシュなテストデータによる評価値の間の差が大きくなる。このことを利用することで、過剰適合を検出することが可能となることを補足しておく。

- Dwork らの報告[25]

Dwork らは、古くから知られている Freedman’s paradox の実験[26]に触発されて次のような 2 クラス分類の実験を行った。

方法：まず、正規乱数を用いて 10,000 個の特徴量を持つ 10,000 個の dataset を 3 セット発生させ、それぞれを training、holdout[3]、test の 3 つの独立した dataset とした[4]。ここで、各データには 2 クラスの内のいずれかのラベルをランダムに割り当てているため、正しい分類精度は 50%になるはずの実験である。次に、線形の分類器の学習をこれらのデータを用いて行うが、論文では、holdout dataset による評価結果を繰り返し利用しながら学習する方法を採用した。具体的には、training と holdout の dataset を用いて、正解クラスラベルと高い相関を持つ（高い分類精度を持つことが期待される）特徴量を選抜し、それらの中から、training と holdout の相関が同じ符号を持つ特徴量を選択して分類器を設計している。

結果：通常行う training dataset だけを用いた分類器の設計ではなく、holdout dataset も参照しながら分類器の設計を繰り返すことで、holdout dataset による評価値が真の評価値から偏り、選択する特徴量が 500 個になったとき、training dataset と holdout dataset に対する分類精度は 63±0.4%となった。一方、完全にフレッシュな test dataset を用い

[3] 「holdout set」はモデルの複雑さを最適化する確認用集合で validation set とも呼ばれる[30]
[4] 「holdout data」及び「フレッシュなテストデータ」が「test data」に相当する

て評価した場合には 50%付近であった（図 1. A 参照）。このように、本来 50%付近になるはずの分類精度が見かけ上高くなってバイアスが加えられてしまったのは、holdout dataset を繰り返し用いて特徴量を選択したため、holdout dataset に過剰適合した結果である。彼らは、一部の特徴量に正解クラスラベルとの相関が存在する場合についても同様の実験を行い、そこでも同様にバイアスの問題を指摘している（図 2. A 参照）。

解決策：差分プライバシーに基づくアルゴリズム（Thresholdout）

Dwork らはこの問題に対処するために、Thresholdout と呼ばれるアルゴリズムを提案している。このアルゴリズムは、差分プライバシー[27]の研究成果を踏まえて生まれたアルゴリズムである。差分プライバシーの方法に基づいて holdout dataset を利用することで、holdout dataset に過剰適合することを防ぐことを狙っている。

図 3 に具体的なアルゴリズム、図 1. B と図 2. B に、そのアルゴリズムを用いて得られた値を元に再学習した分類器に対して、training、holdout、test の各 dataset で評価した結果を示す。ここで、図 3 のアルゴリズム 2(a)の行の$\varepsilon_{S_h}[\phi]$や$\varepsilon_{S_t}[\phi]$は、holdout と training の dataset を用いてそれぞれ計算されている。この行では、holdout と training の値の差の絶対値が右辺の値より大きい場合には holdout で求めた値にノイズを加えて返し、そうでない場合には、training による値を holdout の値として返す（図 3 の 2(b)の行）。直感的には、holdout と training による値が互いに類似している場合は、ユーザのアクセスは training dataset に限定され、そうでない場合のみ holdout dataset へのアクセスが可能となる。ただし、その holdout dataset による値にはノイズを加える。これにより、holdout dataset を分類器の設計に繰り返し利用した場合にも、holdout dataset への過剰適合を防ぐことが可能となる。このアルゴリズムで得られた値を用いて分類器の再学習を行うことで、図 1. A や図 2. A の holdout（緑）の結果が training（青）に近づいていた問題が、同図 B に示した fresh な test による赤色の結果に近づいて解消されていることが分かる。すなわち、分類器の再学習において同一テストデータを繰り返し用いて評価し、かつ、その評価結果を利用して分類器を設計しても、テストデータに対する過剰適合が生じないことが確認できる。

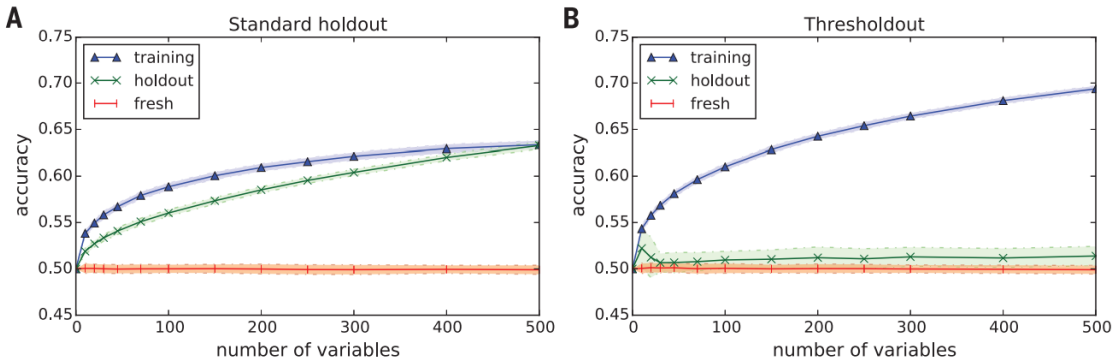


図 1.　特徴量と正解クラスが無相関の場合（文献[25]の Fig.1 より転載）

(A)通常の評価法による結果

(B)提案する Thresholdout を用いた評価結果

縦軸：100 回の施行によるそれぞれの dataset に対する分類精度の平均と標準偏差

横軸：分類器で選択された特徴量の数

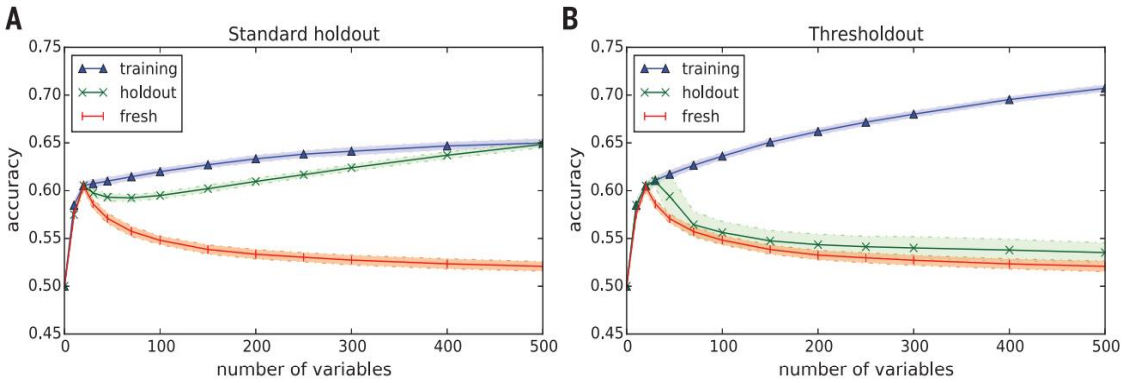


図 2.　一部の特徴量と正解クラスに相関が存在する場合（文献[25]の Fig.2 より転載）

(A)通常の評価法による結果

(B)提案する Thresholdout を用いた評価結果

縦軸：100 回の施行によるそれぞれの dataset に対する分類精度の平均と標準偏差

横軸：分類器で選択された特徴量の数

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 $\mathcal{E}_{S_t}[\phi]$.

図 3.　Thresholdout のアルゴリズム

（文献[25]の supplementary materials の Fig.S1 より転載）

➢　Gossmann らの報告[28]

Gossmann らは、Dwork らの研究に基づいて、SaMD の評価値に AUC を用いた場合のアルゴリズム $\mathrm{Thresholdout}_{\mathrm{AUC}}$ を示している[28]。図 4 にそのアルゴリズムを示したが、図 3 とほぼ同様である。主な違いは、Dwork らの図 3 では holdout と training の dataset による評価値として特徴量と正解ラベルとの相関などを用いていたのに対して、

Gossmann らの図 4 は AUC を用いている点である。Gossmann らは、リスクのある再学習法として、分類に用いる特徴量を test[5]と training の dataset により評価した AUC を用いて決定する方法を例として取り上げ、$\text{Thresholdout}_{\text{AUC}}$ を用いることでバイアスが小さくなることを人工データと実データに対して示している。

図 5 は、頭部 CT 像からの脳出血の有無を分類する問題に適用した結果である。これから、同一テストデータを繰り返し用いて特徴選択を行うことで存在していたバイアスが、$\text{Thresholdout}_{\text{AUC}}$ を使うことで小さくなっていることが分かる。

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

図 4　$\text{Thresholdout}_{\text{AUC}}$ のアルゴリズム（文献[28]の Algorithm 3.1 より転載）

[5]Gossmann らのアルゴリズムの「test dataset」、Dwork らのアルゴリズムの「holdout set」が「同一テストデータ」に相当する

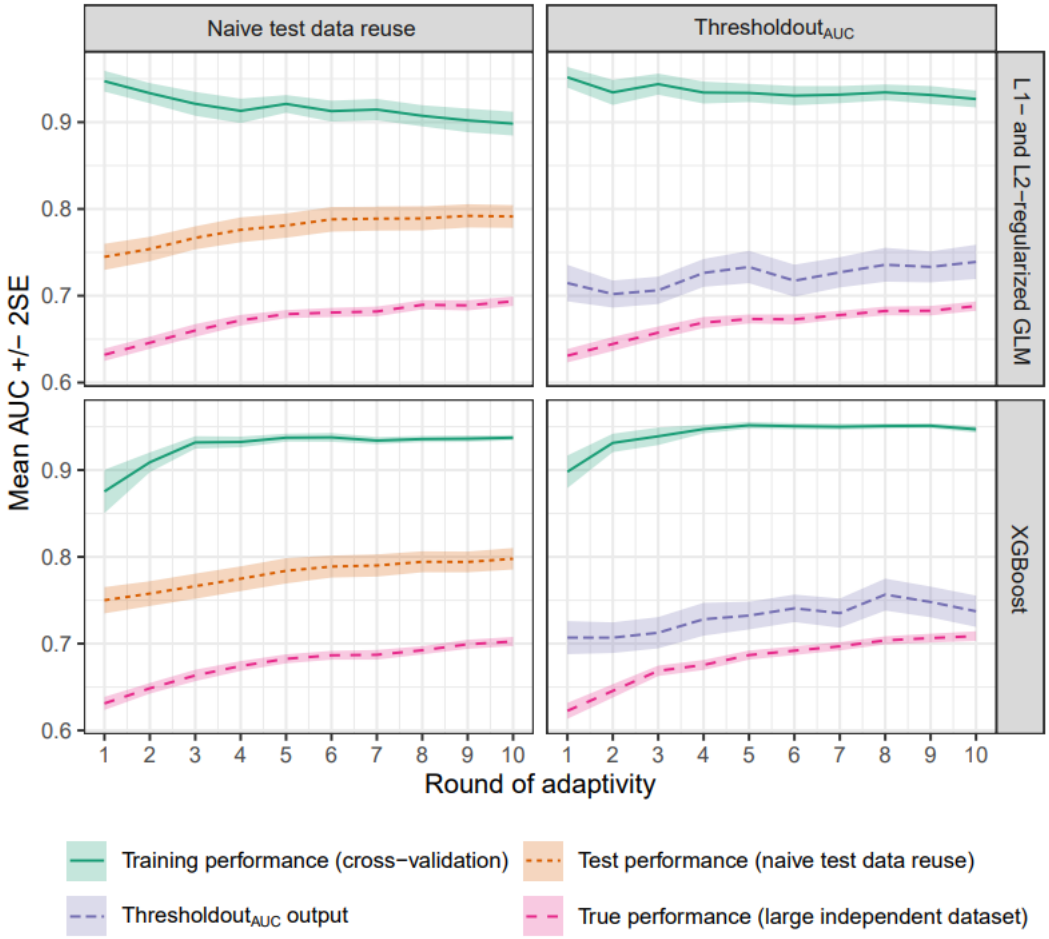


図 5.　頭部 CT 像上の脳出血の有無の分類問題における平均 AUC と標準偏差の推移（文献[28]の Fig.3(a)より転載）

横軸　　：特徴選択の回数

緑　　：学習データにより評価した性能

オレンジ　　：同一テストデータを繰り返し特徴選択に利用した性能
（Thresholdout$_{AUC}$ を用いなかった場合）

紫　　：同一テストデータを繰り返し特徴選択に利用した性能
（Thresholdout$_{AUC}$ を用いた場合）

赤　　：フレッシュな大きい独立したデータを用いて評価した性能

（上段は一般化線形モデル、下段は XGBoost の結果）

なお、Thresholdout アルゴリズムにも課題が残されていることを指摘しておく。上記のようにバイアス削減の効果は確認できたが、図 3 や図 4 のアルゴリズムには、バイアスの削減の程度に影響を与える、ノイズのパラメータのτやσ、及び、テストデータへのアクセスの可否を決定する閾値$\hat{T}$ がある。これらの最適値は問題ごとに異なり、バイアス削減の効果を最大限発揮させるためには、テストデータへの過剰適合によるバイアス

を評価しながら実験的に最適化する必要がある。また、その際のバイアスの正確な評価には、フレッシュなテストデータによる真の評価値が必要となるなど、アルゴリズムの適切な利用のためにはこのような課題が残されている。

４－４．Kaggle における同一テストデータによる評価とそのリスクの評価

近年、画像などを対象とした機械学習のコンペティションが盛んに行われている。Kaggle によるプラットフォームはその最大のものであり[29]、2019 年当時、1,461 のコンペティションが開催された[30]。コンペティションでは、2 つの異なるテストデータによるスコアが用いられる。リーダーボードに公開される public score と呼ばれるスコアと、最後に順位を決定する private score と呼ばれるスコアである。参加者は、直前まで公開されている public score のみを見て繰り返しプログラムを改良することが可能である[6]。これは、同一テストデータによる評価と改良を繰り返すことに相当する。また、フレッシュなテストデータによる評価は、private score に相当すると考えられる。ただし、参加者はすべてのテストデータを受け取り、それに対する予測を行っているため、厳密な意味では完全にはフレッシュではない。しかし、テストデータの正解ラベルは与えられておらず、順番はシャッフルされ、どれが public でどれが private であるかが分からないため、これらを意図的に明らかにする工夫をしなければ、private による評価はフレッシュなテストデータによる評価に近くなると期待できる。

Roelofs らは[29]、Kaggle のコンペティションの中から、参加プログラム（submissions）数などの条件を使って 120 の分類タスクのコンペティションを選び、public と private の score を詳細に比較した。まずは、全参加プログラムの score を用いた解析、次に、public score が上位 10%のプログラムに注目して解析を行っている。さらに、いくつかの定量的評価も試みている。図 6 及び図 7 は、最も参加プログラム数が多い 4 つのコンペティション（表 1）についての public と private の score の関係を示している。

表 1.　参加プログラム数がもっとも多い 4 つのコンペティション（文献[29]の Table 1 より転載）

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}：public test データ

$n_{private}$：private データの数

[6] 1 日の提出回数や最終提出回数に制限があることがある

図 6 からは、public と private の score が 45 度の直線に沿って分布しており、public への明らかな過剰適合は見られない。しかし、上位 10%の score をプロットした図 7 からは、例えば competition 3788 では 45 度の直線より下に分布し、public の方が性能は優れており、public に対する過剰適合の可能性が示唆されている。しかし、両者の差は大きくなく、実際、縦軸の目盛りからは差は 1pt 程度と小さいことが分かる。

120 のコンペティションについて同様の検討をした結果、public に対する過剰適合が強く疑われた例は、public と private の分割法が不適切で分割後の二つのデータが独立同一分布（independent and identically distributed; IID）であるという仮定が成り立っていない場合や、public や private のデータ数が少ない場合であり、それを除けば Kaggle の中では public と private の score は同程度であり、過剰適合の影響は限定的であると結論づけている。その他、興味深い点としては、public と private の score の差とテストデータ数との関係について調べており、データ数が多いほど、両者の差が小さくなることが示されている（図 8 参照）。

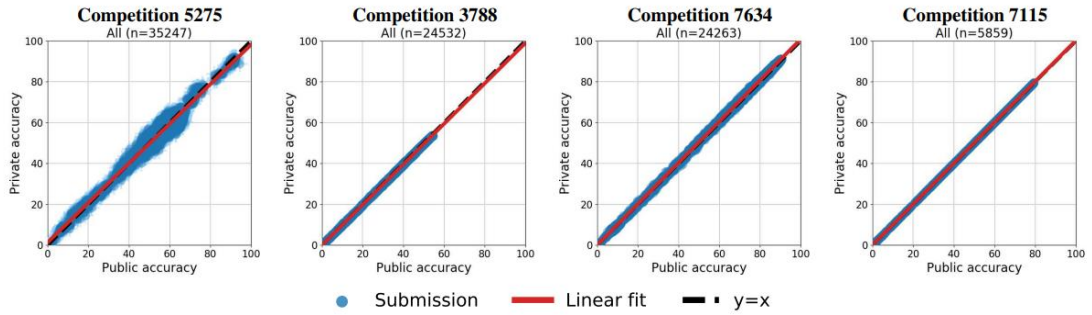


図 6.　参加プログラム数が多い 4 つのコンペティションにおける全参加プログラムの public と private の score の関係（文献[29]の Fig.1 より転載）

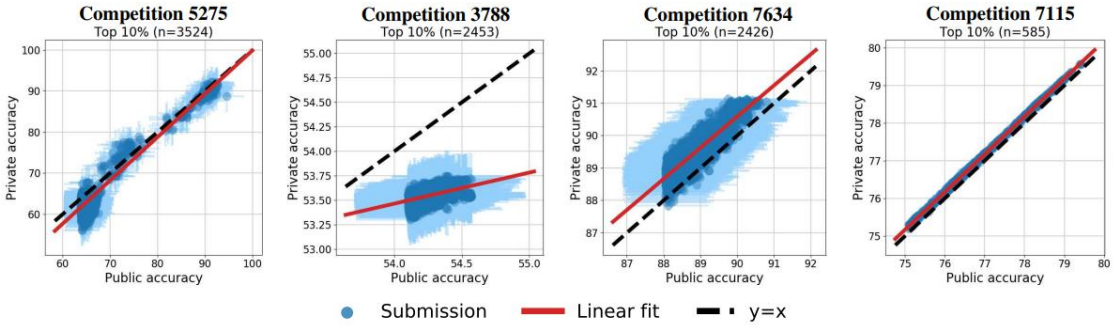


図 7.　参加プログラム数が多い 4 つのコンペティションにおける上位 10%のプログラムの public と private の score の関係（文献[29]の Fig.2 より転載）

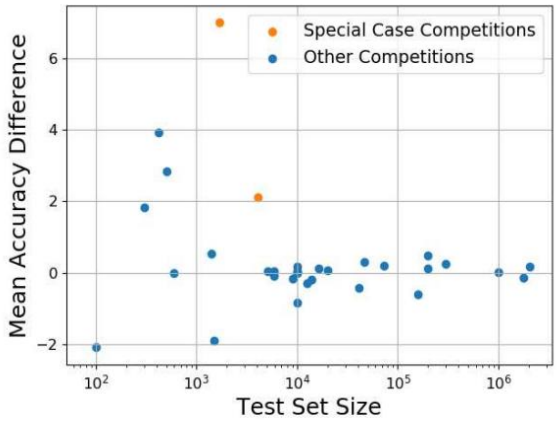
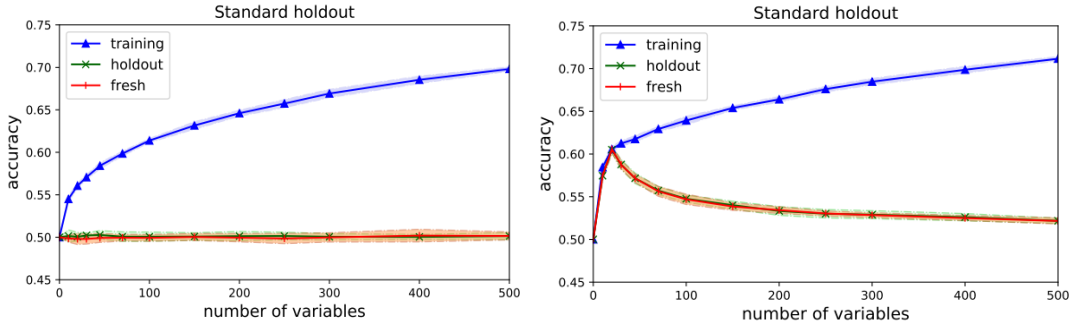


図 8.　平均精度差とテストデータサイズ（public と private の合計）の関係
参加プログラム数が 1,000 以上の 34 のコンペティションについて調べた結果
（文献[29]の Fig.5 より転載）

➢　Dwork らのデータを用いた追加シミュレーション

Kaggle のコンペティションのデータを用いてテストデータへの過剰適合によるバイアスを評価した結果では、ほとんどのコンペティションで同一テストデータによる評価が繰り返し行われているにもかかわらず、バイアス（public と private の score の差）は小さいと結論付けられていた。この結論からは、Dwork らが取り上げた「同一テストデータによる評価結果を繰り返し利用しながら分類器を再学習する方法」を意図的に採用しなければ、バイアスが含まれる可能性は小さくなることが示唆される。

そこで、Dwork ら[25]の図 1 と図 2 の実験に対してこのことを確認するために、同一テストデータ（holdout dataset）による評価結果を再学習に利用しなかった場合について、追加でシミュレーションを実施した（図 9 参照）。どちらのグラフも Thresholdout を利用していない結果であるが、holdout（緑）と fresh（赤）は、ほぼ等しい性能を示している。このことから、テストデータの評価結果を再学習に利用する不適切な方法を意図的に採用しなければ、テストデータへの過剰適合によるバイアスはほとんど生じないと考えられる。SaMD 開発時には、cross validation 法のように、手持ちのデータをランダムに学習用、検証用及びテスト用に分割し、学習と性能評価を繰り返し行っている。このような場合にも、テスト用データによる評価結果を意図的に学習に利用しなければ、Dwork らが指摘したバイアスはほとんど生じないと考えられる。



(A) (B)

図 9. Dwork らの図 1 と図 2 の実験において holdout dataset による評価結果を再学習に利用しなかった場合の結果

(A) 特徴量と正解クラスが無相関の場合

(B) 一部の特徴量と正解クラスに相関が存在する場合

いずれも Thresholdout を利用していない結果

縦軸：100 回の施行によるそれぞれの dataset に対する分類精度の平均と標準偏差

横軸：分類器で選択された特徴量の数

この他、無意識のうちに同一テストデータによる評価結果を分類器の設計に反映をさせてしまう可能性についても指摘をしておく。例えば、開発者が同一テストデータによる評価結果を見た後で、開発者が持っている分類器の設計に関する暗黙知を無意識的に利用する場合である。上記４－３．で示した評価結果を直接利用する場合と比べてバイアスは大きくないかもしれないが、これもリスクの一つとなる。

４－５．SaMD 開発におけるリスクの例

ここまでに紹介した研究論文やそれに対する考察及び最近の PMDA での審査の事例[31]等に基づいて、以下では、現時点で確認できる SaMD 開発におけるリスクが高い行為とリスクが低い行為について例示する。なお、リスクの高低の程度は問題によって異なること、リスクが高い行為と低い行為を同時に実施した場合にどちらの影響が優位になるかは分からないこと、以下の例はすべての場合を網羅していないことなどに注意されたい。また、SaMD 開発中にリスクが高い行為を 1 回でも実施した場合には、開発された SaMD の評価結果にはバイアスが含まれること、行為の回数が増えるほどバイアスがより大きくなる可能性が高いことにも注意されたい。

SaMD 開発におけるリスクが高い行為

- 同一テストデータによる評価結果を SaMD の設計に意図的に利用する行為
  - テストデータによる評価結果を用いて SaMD のパラメータの全体や一部を決定すること

- アルゴリズム、パラメータ、モデルの構造などが異なる複数の SaMD の性能を同一テストデータを利用して比較し、それに基づいて SaMD を選択すること
- 処理に失敗をしたテストデータに注目し、それに類似したデータのみを学習用データとして追加して SaMD を設計すること

➢ 同一テストデータによる評価結果を SaMD の設計に無意識に利用する行為
（同一テストデータによる評価結果を開発者が詳細に解析することで、無意識のうちに開発者の持つ暗黙知が SaMD の設計に反映される可能性がある）

SaMD 開発におけるリスクが低い行為

➢ 同一テストデータへの開発者のアクセス制限などをする行為
- 開発者によるテストデータへのアクセスを物理的に制限し、テストデータによる評価結果が意図的な製品改良に結び付かない開発環境を用意すること[31]。
- やむを得ない理由によりテストデータによる評価結果を再利用する場合には、Dwork らが提案した差分プライバシーに基づく方法[25, 28]を利用すること。具体的にはテストデータへのアクセスを制限し、テストデータによる評価値が必要な場合には、ノイズを加えた後で利用する。ただしこの方法は、加えるノイズのパラメータやアクセスの可否を決定する閾値を適切に設定することが重要となる。

➢ テストデータのサイズを大きくすることはリスクの低減につながる
（例えば、Kaggle のメタ解析の結果（図 8）からは、特別なケースを除いて約 5,000 以上のサイズではバイアスが大きく減少することが示されている）

ここまでに示したリスクは、完全にゼロにすることは難しい。そのために開発側や審査側は上述のリスクについて正しく理解するとともに、リスクを低減する工夫を重ね、それらについて科学的・客観的に説明することが重要である。

リスクにより生じたテストデータへの過剰適合の検出や、検出後の対策も重要となる。検出法としては、例えば、再学習の評価のために繰り返し利用するテストデータセットと、過剰適合を評価するためのテストデータセットを用意し、両者の評価値の差が統計的に有意に大きくなった場合に過剰適合を疑う方法が考えられる。この際、2 つのデータセットが独立同一分布であることや、評価値が統計的に信頼できる程度にデータ数が十分であることに注意が必要である。また、審査側は過剰適合を評価するためのテストデータセットの情報（正解ラベルなども含む）を開発者側に漏れないようにすることも重要となる。

過剰適合を起こした場合には原因を突き止め、リスクが低い開発方法を採用して SaMD 開発を継続する、あるいは、SaMD の利用停止も含む厳しい措置によって問題の拡大を未然に防ぐなどの対策が必要となろう。今後も、これらの議論を深めることが重要である。

５．医用画像（放射線科領域の画像、超音波画像を主体とした医用画像）を用いた深層学習 AI 開発におけるバイアス

AI 研究におけるバイアスに関わる具体的な問題として、医用画像を用いた深層学習 AI 開発におけるバイアスに関する公表論文を例に、以下に関連する最近の議論を紹介する。

- 研究開発方法の問題
- 対象とする医用画像のバイアスの問題

５－１．研究開発方法の問題[32]

画像診断に関する前向き深層学習（deep learning）研究や無作為化試験は少なく、非無作為化試験のほとんどは前向き研究ではなく、バイアスのリスクが高く、既存の報告基準から逸脱していることが、示されている。また、多くの研究は、データやコード（データの前処置とモデル化に使用）を利用できず、対象群の専門医数が少ないことも示されている。

対象：Medline、Embase、Cochrane Central Register of Controlled Trials 及び 2010 年から 2019 年 6 月までに WHO（世界保健機関）に登録された臨床試験に関する文献データから、医用画像処理における深層学習アルゴリズムの性能を 1 人以上の専門臨床医のグループと比較する無作為化試験登録及び非無作為化研究を抽出し分析を行った。医用画像処理では、深層学習研究への関心が高まっており、特に深層学習における CNN の主な際立った特徴は、CNN に生データが供給されると、パターン認識に必要な独自の表現が得られることである。このアルゴリズムは、使用する特徴を人間が指示するのではなく、分類にとって重要な画像の特徴を学習する。分析対象とした研究報告では、既存の疾患の絶対リスクを予測するため、あるいは診断結果群（疾患又は非疾患等）に分類するために医用画像を使用することを目的とする研究である。

方法：レビュー方法として、無作為化研究では CONSORT（臨床試験の報告に関する統合基準）を使用し、非無作為化研究では TRIPOD（個人の診断後または診断のための多変量予測モデルの開発、検証、更新に関する報告の質の改善を目的としたチェック手法）を使用して、臨床研究成果の報告基準への遵守が満たされているかどうかを評価した。バイアスのリスクの評価には、ランダム化研究である場合には Cochrane risk of bias tool を使用し、非ランダム化研究には PROBAST（prediction model risk of bias assessment tool）を使用して評価している。

結果：深層学習の無作為化臨床試験は 10 件であり、そのうち 2 件は公表されている（盲検化が行われていないという低いバイアスのリスクを有するものの、臨床研究結果の報告に係る基準への順守は高いレベルで行われていた）。残る 8 件は進行中の臨床試験であった。非ランダム化臨床試験 81 件のうち、9 件だけが前向きであり、6 件だけが実際の現実の臨床環境で試験が行われていた。AI モデルの出力結果と比較する対

象データの作成に従事する専門家の人数の中央値はわずか 4 人であった（四分位範囲 2〜9）。研究に使用したデータセットとコードへの完全なアクセスができる例は著しく限定的なものであった（それぞれ研究の 95%と 93%でアクセスが不可能であった）。全体的なバイアスのリスクは 81 件の研究のうち 58 件で高く、臨床試験の報告として順守すべき基準の順守は適切とは言えないものであった（29 件の TRIPOD 項目のうち 12 件で順守率が 50%未満であった）。81 件の研究のうち 61 件が、AI のパフォーマンスが臨床医のパフォーマンスと同等又はそれ以上であると述べていた。81 件の研究のうち 31 件（38%）のみが、さらなる前向き研究又は試験が必要と述べていた。

結論：これらの既存研究報告の分析より、医療用画像処理分野では、深層学習に関する前向き研究や無作為化試験はほとんど行われておらず、またほとんどの非無作為化試験は前向きではなく、バイアスのリスクが高く、臨床研究報告が満たすべきものとして存在する報告基準から逸脱していた。ほとんどの研究では、データとコードの可用性が不足しており、比較対象のデータを作成する人間のグループはしばしば小規模であった。今後の研究では、バイアスのリスクを減らし、現実世界の臨床関連性を高め、報告の手法とその透明性を改善し、結論を適切に調整する必要があると結論づけている。

一般に臨床試験に使用する医用画像には、人種、性別、体格、疾患分布などバイアスが存在するのが通例である。このバイアスリスクを低下させるために、無作為化（比較）試験や前向き研究が行われていることは言うまでもない。この論文[32]は、臨床試験という観点から深層学習研究開発を分析した結果であり、これまで深層学習を用いた医用画像処理に関する研究報告は、現在の研究内容が臨床試験としては必ずしも質の高くないものが多いことを述べている。この論文が求めるように、一般の臨床試験で求められる前向き研究や無作為化試験をそのまま深層学習モデルの研究開発に当てはめることは、現在の同技術の研究開発段階を考えると必ずしも適切であるとはいえないとも思われる。

AI 医療機器の認可の障害とならないよう、今後の関連技術の研究開発動向を考慮しつつ、より適切な臨床評価の在り方については議論していかなければならないと考えられる。しかし当然のことであるが科学的にバイアスのリスクを減らすための努力は必要であり、さらに市販後学習により性能改良が行われたことを評価するために、同様に前向き研究や無作為化試験が市販後に必要とされるのかについても引き続き議論が必要であろう。

５－２．対象とする医用画像のバイアスの問題[33]

開発者が意識することが難しい、学習データ、テストデータに公開されている別目的で開発された医用画像データに潜在的に存在するバイアスの問題に触れている。

公開データベースは現在の深層学習時代の重要なデータソースとなっているが、これ

はその公開データベース開発された目的とは異なる、「オフラベル（適応外）」で使用されることにもなる。あるタスクのために公開されたデータが、別のタスクのアルゴリズムをトレーニングするために使用されることが広く行われている。この論文は一般的な深層学習を用いた解析システムの研究開発において行われている慣行が偏った過度に楽観的な結果につながる可能性があることを指摘することを目的としている。

方法 : MRI（核磁気共鳴画像診断）信号から MRI 画像を再構成する逆問題ソルバーについてこの現象を実証し、公開データベースに典型的な 2 つの処理パイプラインについて説明し、磁気共鳴イメージング測定データからの画像再構成をする逆問題ソルバーを研究対象として、当該分野で確立された 3 つのアルゴリズム（圧縮センシング、辞書学習、深層学習）への影響が検討されている。これによりそのバイアスを持つ性能が隠れたデータ処理パイプラインに起因することを示す。

結果 : すべてのアルゴリズムが一見適切なデータと考えられる学習データで訓練された場合、体系的に偏った結果をもたらすことを示している。正規化された二乗平均平方根（Root Mean Square; RMS）誤差は、データ処理の程度に応じて一貫して改善され、場合によっては 25〜48%の人為的な改善が見られた。この現象は広く知られていないため、偏った結果が最新技術として公開されることがある。公開データベースは機械学習の研究にとって重要なリソースであるが、機械学習への利用可能性が高まるにつれて、あるタスクのために公開されたデータが別のタスクに使用される「適応外」の使用につながることがある。この研究は、このような適応外の使用が、偏った、過度に楽観的な機械学習アルゴリズムの性能評価につながる可能性があることを明らかにしている。

この結果に基づけば、MRI の測定データから MRI を再構成する逆問題ソルバーの機械学習を行う場合には、学習データとして使用する MRI 信号と再構成画像のペアで表現される学習データセットに関して学習に使用した画像が、どのベンダー、画像フィルター、撮像条件で生成されたものであるかといった詳細な撮影パラメータなどを明らかにすることが望ましい。しかし実際に利用可能な公開データベースに対してこれらを全て明らかにすることは困難と思われる。少なくとも real world での性能評価を適切に行うためには、少なくともテスト画像に関しては、それら MRI の各種パラメータを明らかにする必要がある。この論文からは、実際の様々な市場データでの成績は、研究開発のテスト成績より劣ってしまうことは十分にあり得るということを示唆している。このようなことは MRI のみならず、CT や超音波画像でも十分起こり得る現象である。この制約を実際に適応する場合に、どの程度の許容度にするかは議論が残る課題である。

研究開発時に、そのテスト結果があまりに良好である場合は、医用画像 AI 研究開発に使用した医用画像が、あまりに特殊な撮影条件の画像や、過剰な画像処理による一般的ではない画像である可能性も考慮すべきである。わかりやすい例を挙げれば、3.0T の

MRI で開発された医用画像 AI を 1.5T の MRI 画像で使用する場合などである。また 64 列の CT 画像で作成された AI を 16 列の CT 画像に使用する場合や Adaptive Filter を使った超音波画像で開発された医用画像 AI を Fundamental US 画像で使用する場合などが考えられる。

しかし、医療機器認可の審査の際に、これらの詳細なパラメータに制限をかけて適応を狭めすぎることは、結果として AI の臨床利用の障害となり得ることを考慮すべきであり、リスクベネフィットバランスの観点から適切な設定がなされるべきであろう。

６．物理モデル・シミュレーションによる学習データ構築の現状と課題

数値シミュレーションと機械学習を併用する研究も進められており、近い将来に数値シミュレーションを開発過程で利用した機械学習による医療機器プログラム、又は数値シミュレーションと機械学習を併用した医療機器プログラムが登場することも予想される。本項では、執筆時点で機械学習における数値シミュレーションの利用において留意すべき点を考察した。

数値シミュレーションは注意深く用いることで、実計測のデータで課題となるバイアスを含めてコントロールすることが可能である一方、数値シミュレーション特有の制限があることに留意する。それらに関しては、科学委員会の報告書が参考になる[3]。本項での用語は当該報告書に準拠した。

６－１．機械学習における数値シミュレーションの利用

現時点では、機械学習における数値シミュレーションの主な利用法としては以下がある。

1. 数値的に合成したデータを用いた学習
   例として、McGill 大学「BrainWeb MRI Simulator」がある。
2. 学習データに数値シミュレーションで摂動を加えた学習
   画像の機械学習においては、以下の例の加工を行った画像を学習させる。検出対象の大きさ、位置、画素値等への依存性をなくすために必要なデータ・オーギュメンテーションとして一般的である。なお、摂動を加えて得たデータを評価データとして用いることは、その必要性を含めて現段階では位置付けと解釈が定まっていない。
   - 変形（回転、平行移動、拡大等の剛体変換、Free Form Deformation 等の非剛体変換）
   - 濃淡値変化（濃度値のガンマ変換、ランダムノイズ付与等）
   - 潜在空間の特徴ベクトルにノイズを付与することで、画像の主要な特徴を維持しつつ細部の特徴を変化させる生成モデルによる加工
3. 数値シミュレーションの結果をラベル（正解）とする学習
   特に GAN における利用例がある[34, 35]。医療分野では、Nguyen らは顔面神経麻痺、顔面再建等の治療後の予後予測、リハビリテーション等の目的に用いる「表情筋を動かして表情を生成するモデル」を GAN と有限要素モデルの組み合わせにより構築した。個々の表情筋の弛緩緊張を表すモデルにより笑顔と左右対称の表情を作るタスクを学習させた[34]。表情筋の弛緩緊張から顔の様子（表情）を計算する有限要素モデルと、既知の表情データベースを比較して生成器（generator）と識別器（discriminator）と学習させる。この例では顔面神経麻痺患者等の表情筋の状態に関するデータを用いずに学習を行わせている。

６－２．数値シミュレーションを機械学習に用いるメリット

現在のところ、医療機器の分野で応用可能な数値モデルの多くは物理モデルであり、それ以外の生理学的なモデルは種類も、適用可能な範囲も限定的である。機械学習と組み合わせることで数値シミュレーションのみでは扱えなかった問題を扱うことが可能となる。

上記６－１．に示すいずれの利用方法においても、数値シミュレーション部分は順問題として記述可能な場合、静定問題となる場合が多いと考えられる。順問題であれば、数値計算に伴う技術的問題が少なく、解の安定性も期待でき、複数の状態方程式を連成させて解くといった複雑な処理も不要又は独立して扱うことができる可能性がある。

医療の場合、必要なパラメータが患者から得られない又は代表値しか得られないといった場合が多い。数値シミュレーションを、ネットワークを鍛える目的に限定して使用する方法は、パラメータに関する曖昧さの影響をネットワーク側に寄せることで、開発の柔軟性を高める可能性がある。

数値シミュレーションを含む生成データを用いることは、希少疾病、脆弱な被験者群からの同意取得の倫理的課題、または個人情報取扱い上の課題の解決、必要な実データ数の減少、実データ収集に要する期間の短縮につながる可能性、また研究開発用のデータセットの共有を促進する可能性がある[36]。

６－３．数値シミュレーションを機械学習に用いる際の留意事項

数値シミュレーションと機械学習が相互に影響し合うため、実装過程と verification and validation（V&V）の過程が複雑化する。例えば意図しない挙動が起きた場合に、どちら側に問題があるのか切り分けて検証しなければならない。

data driven model による数値計算は ASME V&V 40 の適用範囲外となっている。機械学習モデルを含む全体のプロセスに、また医療機器におけるシミュレーションの信憑性(credibility）に関する ASME V&V40 を適用すること、uncertainty quantification の導出は未解決の課題である[37]。

冒頭でデータへのバイアスをコントロールできる可能性があると述べたが、一方で数値シミュレーションが表現可能な事象の制限を受ける。数値シミュレーションでは、非線形性、特異性の高い系、過渡応答や非定常的な系の挙動の扱いの難度が高くなる。例えば、破壊現象を扱うにはさまざまな単純化（近似）が必要となる。そのような近似の妥当性を吟味する必要がある。また、数値シミュレーションで病変、病態、外傷等をモデル化した場合、そのモデルの妥当性検証が必要な場合がある。

形状、物性等の数値シミュレーションに必要なパラメータを変更しながら多数のデータを生成してデータ・オーギュメンテーションを行う場合、パラメータの変更、その変動幅が数値シミュレーションの結果の持つ不確かさ、演算効率に及ぼす影響を検討する必要がある。例えば有限要素法において、メッシュ形状と細かさは計算結果の不確かさ

と演算時間を左右する。メッシュの寸法を局所的に変えて極端に細長いメッシュにすると、計算結果の不確かさを増大させる。メッシュを全体的に縮小させて細かすぎるメッシュにすると、演算時間を無駄に増大させる。

メッシングをやり直すことでこれらの影響を排除することができるが、演算時間が長くなるだけでなく、やり直したメッシュの妥当性評価も含めて必要になる。

一方、データ・オーギュメンテーションに用いるデータは、個々のデータの妥当性検証を全例について実施する必要性はないと考えられる。例えば医用画像のデータ・オーギュメンテーションで慣行されるアフィン変換は、画像中の関心領域の大きさ、向き、形状について一般化することで特定の大きさ、向き、形状への過学習を防ぐ役割がある。この場合は画像データをアフィン変換することの医学的妥当性を求める必要はない。

６－４．数値シミュレーションと機械学習の今後

数値シミュレーションと機械学習の住み分け

これらに鑑みると、数値シミュレーションで生成したデータを機械学習に用いる場合は患者から得たデータでは不足する健常状態の学習と言った、数値シミュレーションと臨床でのデータ取得の得意不得意を相互に補う利用方法が望ましい。表情を生成するモデルの例[34]では、表情筋の緊張弛緩を入力とする数値モデルにより顔表面の表情をシミュレートさせており、数値シミュレーションにその得意とする部分を分担させていると言える。

数値シミュレーションによる評価データ生成

数値シミュレーションで生成したデータの機械学習への適用対象としては、学習データ（バリデーションデータを含む）が一般的である。一方、評価データ（機械学習の最終モデルの性能を評価するために使用されるデータ）として数値シミュレーションにより生成したデータを用いるには、克服すべき課題が存在する。

- 数値シミュレーションによる学習データ

  数値シミュレーションの信頼性（reliability）、信憑性（credibility）は、基盤となる数値シミュレーションで用いるモデルの成り立ち、シミュレーションに用いるパラメータ等の根拠となる「エビデンス」によって左右される（コンピューターシミュレーションを活用した医療機器ソフトウェアの審査の考え方に関する専門部会報告書　第６章[3]）。

  演繹的に導出されたモデル（対象の患者群から得られたデータに基づく数値シミュレーション）と比較すると、実験式的なモデル（健常者や動物での計測で得られたデータに基づく数値シミュレーション）では信憑性が及ばないとされる。

- 数値シミュレーションによる評価データ

  現段階では評価データは実測データであるべきとの考えが主流である。一方で、機械学習の評価は end-to-end で行う治験（あるいは「追加的な侵襲・介入を伴わな

い既存の医用画像データ等を用いた診断用医療機器の性能評価試験の取扱いについて」（令和 3 年 9 月 29 日付け薬生機審発 0929 第 1 号厚生労働省医薬・生活衛生局医療機器審査管理課長通知）で述べる性能評価試験）ばかりではなく、特定の性能項目の評価を行うことも考えられる（例：「画像取得時に付随する体動への耐性評価」参照）。その場合、その特定の性能項目に着目して数値シミュレーションを設計、実施し、数値シミュレーションの結果の妥当性を吟味することで、機械学習の性能評価に数値シミュレーションの結果を用いることも想定される。具体的には ASTM V&V 40 における context of use と validation を機械学習で評価したい項目と関連づけて論じる必要がある。

- 人為的に加工又は生成した評価データの位置付け

  数値シミュレーションの結果には抽象化に伴う捨象と曖昧さが存在し、医学では観測事実に並ぶものとして受け容れられていない。経験的には医療機器の開発と評価においては、位置付けと解釈が定まらない要素が増えるほど、全体的な評価プロセスが複雑化して難化する。機械学習の医療機器への応用も発展途上にあり、位置付けと解釈は定まっていない。数値シミュレーションの結果を機械学習モデルの評価に用いることは、医療機器としての機械学習モデルの評価を更に難しくする可能性が高い。一方、工学では適切に得られたシミュレーションの観察事実を補足するものとして利用することがある。医療機器の評価では数値シミュレーションを受け容れる過渡期にあると言える。

数値シミュレーションによるデータ利用のシナリオ例（体動への耐性の学習と評価）

（以下は本報告書の説明のために創作したシナリオであり、実現性は検証されておらず、また推奨を意図するものではない）

体動による画像の歪み（モーションアーチファクト）の影響を受けないことを標榜しようとする画像診断支援システム（CAD）を想定する。これを標榜するには、そのような画像を収集してモーションアーチファクトが存在しても正しい出力が得られることを検証する必要がある。しかし実臨床では体動をなくすように指導するのでそのような画像を集めるのは容易でない。評価目的で意図的に体動させながら撮影することは可能であるが、多数集めることはコスト的にも、場合によっては倫理的にも適切でない場合がある（例：その目的で X 線被曝させるのは正当化が難しい）。

そこで、数値的に体動をシミュレートした画像を生成して、これを学習と評価で用いる。以下の方法で数値シミュレーションを実施、V&V する。

- Context of use：肺の CT 画像から病変部を検出する CAD で、体動の影響を受けにくいことを示すためのデータを機械学習モデルの学習と評価で用いる。
- アルゴリズム：臨床で得られた静止状態の画像を入力として、X 線検出器で計測

される吸収線量を算出、さらに CT アルゴリズムで画像再構成する。吸収線量の計算のためボクセルデータから再サンプリングする際に、呼吸動を加える。

- Verification：実験用 CT 装置を用いてファントムを動かして得た画像と、このアルゴリズムで得た画像を比較して、検出器の計測値、描出される画像の形状と CT 値の偏差が設定許容値の範囲内であることを確認した。
- Validation:臨床用 CT 装置で同じファントムを動かして得た画像と比較、また臨床的に生じ得る体動によるアーチファクトについて放射線医が評価した。その結果、一定の条件の場合には、シミュレーションで生成したアーチファクトが臨床用 CT 装置内でファントムを動かした場合に生じるものよりも過剰であることが判明し、評価した医師もそのように判定した。

  原因としては、臨床用 CT 装置では体動を想定した補正が行われていると推察した。なお補正方法は公開されておらず、これをリバースエンジニアリングしてシミュレーションに加えることは困難である。

結論：シミュレーションで生成した画像を学習と評価に用いる際には、過剰なアーチファクトを無視するように学習したモデルが他の重要な疾病に関する特徴を無視しないことを別途確認する必要がある。

７．現在までに構築されたデータベースの概要と留意すべき課題

個別データベースの概要として、現在までに本邦において様々な医療画像においてデータベースを構築する試みがなされてきた。AMED がサポートした医用画像の前向き収集を行った代表的な 4 つのデータベース（手術動画、病理デジタル画像、心電図、消化器内視鏡）についてそれぞれの概要と課題について、以下の観点から整理した。

1. データベース構築の目的
2. 収集データの種類・規格
3. 現状の収集データ規模
4. 紐づけられる医療情報の種類
5. 異なる個人、地域、医療機関、国、時期におけるデータ収集の必要性
6. 人の手技に対する依存度の有無
7. 撮影・記録に関わるデバイスの多様性
8. 使用デバイスの多様性
9. 主なアノテーション手法・教師データの作成方法
10. 想定される利用方法
11. データベースにおける課題

７－１．手術動画データベースの概要

データベース構築の目的

データベースを構築する目的は、求める医用画像によって異なる。消化管内視鏡検査（治療除く）、病理画像、心電図は直接的な診断を目的として記録する一方、手術動画は治療のすべての工程が記録されるデータである。そのため、データベースを構築する目的は、消化管内視鏡検査、病理画像、心電図においては、診断の補助や自動化を目指すためのアノテーションデータとして用いることである。一方、手術動画においては、手術の技術への客観的評価、手技のサポートや解剖構造などの提示を行うためのアノテーションデータとして用いることが多い。教育目的での利用も考慮されるが、使用するデータ量はアノテーションデータと比較すると限定的であると考えられる。

収集データの種類・規格

収集されるデータは、腹腔鏡手術またはロボット支援下手術の手術動画である。本邦ではまず、手術の術式の定型化が進んだ胆嚢摘出術や S 状結腸切除術などがデータベースに収集されてきた。

収集された手術動画における標準規格は存在せず、拡張子、走査形式、画質などは使用するスコープメーカー、録画機器により異なる。

現状の収集データ規模

本邦では AMED のサポートの元で、国立がん研究センター東病院を中心とし構築したデータベースが既存のデータベースの中でも収集数が最も多く 13 術式、4000 例程度に至る（https://www.s-access.ncc.go.jp/）。収集した疾患領域は、大腸（5 術式、36 施設）、胃（3 術式、21 施設）、肝胆膵（3 術式、26 施設）、前立腺（1 術式、17 施設）、合計（学会除く、84 施設）など領域横断的に行われた。収集した医療機関は大学病院に偏らず、市中病院、がんセンターなど広く収集し、術式も腹腔鏡手術だけでなく、ロボット手術や TaTME 手術など、各領域で実際に施行されている術式を網羅した。術後の臨床的なアウトカムとの相関を評価する上で、合併症のある症例も可能な限り収集した。術式にもよるが、例えば、直腸がんに対する低位前方切除術では約 10%程度、膵腫瘍に対す膵体尾部切除術では約 9%の術中または術後合併症のある症例を収集している。

外科医の技術到達度を評価する上で、日本内視鏡外科学会技術認定の有無や医師の内視鏡手術経験年数についての情報を併せて収集し、同時に内視鏡手術の初学者の動画も収集をお願いすることで外科医の技術差が動画の中でどのような差分を生むのかという観点からも評価できるようにした。

紐づけられる医療情報の種類

手術動画に紐づく情報として、患者背景（性別、年齢、BMI、治療歴、悪性腫瘍の場合臨床病期など）、術者情報（医師経験年数や内視鏡技術認定医、ロボット手術のプロクター資格の有無）、機器情報（スコープの種類、ビデオシステムの種類、データの出力方法）、臨床成績（病理結果、術中合併症、術後合併症、再発情報）などがある。

異なる個人、地域、医療機関、国、時期におけるデータ収集の必要性

手術は、患者により腫瘍の深達度や広がりなどが異なるため術式としては同じでも根治性を保つための切除範囲が異なることがある。また手術手技の習熟度は、医師間差及び施設間差が指摘されているため、このようなバイアスを考慮する必要がある。また、外科医や施設の嗜好により異なる使用機器（鉗子、エネルギーデバイス、ロボットデバイス）が使用されることは一般的であり、また経時的に行われる術式のトレンドや使用機器の変遷もあるため、研究開発の目的によっては多様な手術動画の収集が求められることもある。

人の手技に対する依存度の有無

手術動画は外科医の判断にゆだねられた結果であり、定型化されている術式でも患者の状態や術者、施設によって、内視鏡カメラの操作、剥離や視野展開方法、手順が異なる。

撮影・記録に関わるデバイスの多様性

内視鏡メーカー（ロボット手術含む）、録画機器、撮影モード（NBI など）には多様

性が存在し、特に内視鏡メーカーの違いは色調の変化や画面表示の違いをもたらす。また録画機器により動画のファイルを圧縮する方法が異なる。

使用デバイスの多様性

鉗子、吸引管、吻合器、止血剤など内視鏡手術における使用デバイスの種類は多い一方、ロボット手術に関連した企業は限定的でありその使用機器は専用のものが使用される。内視鏡手術において使用されるデバイスは術者だけでなく、助手も使用するため、4～5 本の術具が画面上に出現することも手術動画の特徴である。

主なアノテーション手法・教師データの作成方法

手術工程のアノテーションは Classification により作成される。臓器・術具、電気メスの通電情報などは Classification, Detection, Semantic segmentation により作成される。いずれも医師による作成または医師の監修が必要となる作業である。工程や臓器の境界は詳細に定義づけしても曖昧になってしまうことがあることは課題である。

想定される利用方法

医療機器としては、術者の術中診断補助・ナビゲーションを行うことで合併症を防止するために用いられること、技術評価などにも利用されることが想定される。

手術動画データベースにおける課題

課題の 1 つ目は、内視鏡メーカーや録画機器により動画の規格が統一されていないことである。動画を出力する時点で動画のメタ情報（走査形式、拡張子など）の規格が統一されておらず、統一作業が煩雑である。

課題の 2 つ目は、手術動画は動画のため、体の外が写る場合や音声が含まれている場合、個人情報に配慮する必要があることである。さらに手術中に胆道造影検査などを行い、検査動画も手術動画に保存された場合、ID や患者名などが表示されることがあり、個人情報を含んでいないか動画を見直し慎重に確認する作業が必要である。

７－２．病理デジタル画像データベースの概要

データベース構築の目的

病理診断に関連するデータベースの目的は大きく 2 つある。1 つは教育目的あるいは診断支援目的のためのライブラリーとなるデータベースである。一般的な疾患の典型的な組織画像のデータベース、また希少がん等の稀な組織型の診断支援の目的で構築されるデータベース等がこれに該当する。もう 1 つは病理診断支援のためのプログラム医療機器等の開発のための病理画像データベースであり、プログラム開発目的のためにアノテーションが付与された一般的な病理画像の膨大な症例数が格納されているデータベースである。

収集データの種類・規格

日本国内ではすでに複数の病理組織デジタル画像のデータベースが構築されている。病理デジタル画像は従来、光学顕微鏡に付属したデジタルカメラや 3D カメラ等で撮影した JPEG（Joint Photographic Experts Group）や TIFF（Tag Image File Format）といったフォーマット形式の静止画像が主流であった。一方近年ではスライドガラス標本を 1 枚丸ごとデジタル化し、光学顕微鏡同様に画像の拡大縮小ができるデジタル画像を作成する特殊な病理デジタル画像作成装置「バーチャルスライドスキャナー」が開発され、画像データベースはほぼこの画像に置き換わりつつある。

バーチャルスライドスキャナーにより作成された病理デジタル画像は Whole-Slide Imaging（WSI）と称される。このバーチャルスライドスキャナーは、多数の企業から多くの製品が提供されているが、企業ごとに WSI のフォーマット形式が異なり、規格が統一されておらず、DICOM（Digital Imaging and Communication in Medicine）規格のフォーマット形式は日本で販売されている機器ではごく一部の機種のみである。また、デジタル画像上でも拡大縮小に耐え得るように、Z 軸方向（標本の深度）も加味しながら自動で画像作成を行うために、1 枚の WSI の容量は大きく、約 200MB から大きいものでは 1GB を超えることもある。さらに病理組織画像同様、病理細胞診画像の WSI も存在するが、組織標本の平均的な厚さが約 5μm 程度であるのに対し、細胞診標本の場合はその厚さが 4～5 倍になるため、組織標本よりも Z 軸方向をより多層化して画像作成を行わなくてはならず、1 つの WSI で 5GB 以上になることもしばしばである。

現状の収集データ規模

AI（深層学習）を活用した病理診断支援プログラム開発研究を目的として、一般社団法人日本病理学会（以下病理学会）が AMED の研究支援の下、国立情報学研究所、東京大学、名古屋大学、九州大学等ともに行った「Japan Pathology Artificial Intelligence (AI) Diagnostics Project (JP-AID)」[38]では、主として 16 大学病院と 7 市中病院から病理組織 WSI を収集した。最終的には、約 20 万枚にも及ぶ WSI を収集[39]したが、このうち、12.2 万枚に関してはアーカイブ化が完了しており、病理診断と紐づいた WSI が、現在、試験的に病理学会会員に限定して公開されている。将来的には会員以外の研究者等も活用できるように調整中である。また、教育目的、病理診断支援のためのデータベースとして、「希少がん診断のための病理育成事業（国庫補助金事業）」の下、病理学会が運営している「希少がんデータベース」に、希少がん（脳腫瘍・骨軟部腫瘍・小児腫瘍・リンパ腫・皮膚腫瘍・頭頸部腫瘍・5 大がんの組織学的希少サブタイプ）の WSI（約 1,700 症例分）やそれに関連した疾患のポイント解説、E-ラーニング用 5 択問題等が保管され、病理専門医の更新単位として活用されている。本事業は令和 5 年時点でも継続しており、さらに毎年約 250～300 症例の希少がん WSI をデータベースに追加登録する予定である。なお本データベースは逆字引機能も備えており、疾患名入力でその疾患に関連す

る WSI や解説などが検索、閲覧できるようになっている。現在は病理学会会員のみに公開しているが、将来的には、国内公開さらに国際公開も計画している。

紐づけられる医療情報の種類

WSI に紐づけられる医療情報としては、臓器名、性別、およその年齢、病理診断であるが、中でも臓器名と病理診断はほぼすべての WSI に付帯されている。一方で希少がんでは、病理診断に年齢が非常に重要な因子となることがあり（特に小児腫瘍では、1 歳なのか 2 歳なのかで診断・予後は異なることがある）、年齢がそのまま付与されているデータもあるため、特に個人情報の取り扱いには細心の注意を払う必要がある。なお、現在 WSI の情報として、「どの時期の画像であるか」という情報は紐づけられていない。しかしこの情報の付与は非常に重要である。なぜならば、例えば世界保健機関 WHO が担当している病理診断の国際分類基準は定期的に見直しが行われており、同じ組織でも過去の病理診断と異なる病理診断名に変更されたり、また新たな知見により以前悪性とされていたものが良性になったりすることがあるからである。そのために「どの時期の画像か」の情報を付与する必要があるし、また研究者が WSI を使う際もそのことに留意して使用する必要がある。先に述べた病理学会の希少がんデータベースでも WHO の脳腫瘍病理診断分類が大きく変更となったため、データベースの改変作業が必要となっている。

異なる個人、地域、医療機関、国、時期におけるデータ収集の必要性

現在本邦では、病理標本に関してはスライドガラスをそのまま「データ」として保管している医療機関がほとんどである。しかしスライドガラスは破損するリスクを有し、コピー等もできないため、原本が失われてしまった場合にはデータそのものが消失してしまうことになる。実際に東日本大震災の際には多くのスライドガラスが破損し、多数の医療機関で病理データが消失する事態となった。一方でデジタル画像は、複製等が可能なデータであり、複数の拠点に分散して保管可能である。データ保管の観点からは、病理画像もデジタル画像化して医療機関内及び外部のクラウドサーバ等にデジタルデータを収集して保管することが重要である。

国際的なデータ収集にあたっては、特に病理診断に関しては、国による診断基準の「微妙な差」にも留意が必要である。例えば「大腸ポリープ」に関しては、いわゆる良性の「大腸腺腫」から「大腸癌」に連続して移行するものがあるが、連続する病変であるがゆえ、腺腫と癌の線引きラインが国によって異なっている。子宮頸部の病変なども、医療訴訟などの社会的背景から、米国と日本では診断基準が異なっている。すなわち国際的なデータベースを構築する際には、どの国からのデータなのかの情報も必要であり、その情報も付与したデータを収集する必要がある。

人の手技に対する依存度の有無

病理診断にはヘマトキシリン・エオジン（HE）という色素を用いてのスライドガラス標本（HE 染色標本）が使用されている。しかしこの HE 染色標本の作製工程は、現在でも作製する検査技師のマニュアル作業による工程が多く、使用する染色液や染色時間により施設間で色調や色域が異なり、肉眼的に判別が可能ほどの差がつく場合もある。通常は4～6μm 程度の薄い組織切片をミクロトームという特殊な機器で薄切しスライドガラス標本を作製するのだが、この薄切作業はどこの国でもほぼ検査技師のマニュアル作業で行われている。そのために組織片の厚さが 4～6μm の範囲に収まらず、薄切者間や同じ検査技師でも日によって厚さが異なることがしばしばである。HE 染色工程も標準化がなされておらず、前述の色調や色域の違いとの相乗効果でさらに大きな施設間格差となる。

撮影・記録に関わるデバイスの多様性

撮影に関わるデバイスに関して、WSI 作成のためのバーチャルスライドスキャナーは、日本では 9 社程度であるが、米国では中小のベンチャー企業も含めれば約 80 社以上の企業が製品を販売している。その WSI のフォーマット規格は先述のようにバラバラであり、多くの企業がそれぞれのフォーマット規格に対応したスライドビューワソフトを独自に開発して提供しているのが状況である。これに対して日本ではある程度の技術基準が必要であるとの議論があり、バーチャルスライドスキャナーやモニターなど、ハードウエアに関する技術基準である「病理診断のためのデジタルパソロジーシステム技術基準」が、病理学会、日本デジタルパソロジー研究会、機器を作製・販売している企業により策定され公開されている[40]。

さらに日本では、厚生労働省によるバーチャルスライドスキャナーの医療機器としての薬事承認基準がある。具体的にはクラス I の一般医療機器として、「病理ホールスライド画像保存表示装置」という医療機器名が、クラスIIの管理医療機器として「病理ホールスライド画像診断補助装置」という医療機器名が付されている。前者は、WSI を保存して表示することはできるが、病理診断には用いることができない病理デジタル画像作成装置であり、後者は、WSI をモニター上に移し出して病理診断に使用できる病理デジタル画像作成装置を指しており、医療の現場では厳密に使い分けが求められている。

主なアノテーション手法・教師データの作成方法

WSI へのアノテーション付与、教師付データの作成はまさに人が行う工程である[41, 42]。この場合、課題となる点が 2 点ある。

1 点目が先述のように「病変の連続性」に起因するアノテーション付与の難しさと不確実さである。病理診断では良性から悪性まで連続的に移行するものがある。もちろん HE 染色標本で明らかに良性、または明らかに悪性と決着する症例が多いが、その中間となる病変も少なくない。例えば、ある腫瘍の病理診断では「良性」「悪性」「中間悪性」という分類が存在するが、この場合、この「中間悪性」をどのようにとるかが病理医間

でも差が大きい。病理医 A が良性と判断したものを、病理医 B は中間悪性に分類する場合や、病理医 B が中間悪性としたものを病理医 C が悪性とするといった「観察者間のバイアス」が、教師データ作成時に少なからず存在する。そのため「目合わせ」なるものがアノテーション付与前に必要となる。

2 点目は、教師付データを作成するために膨大な時間を要する点である。1 枚の WSI を細かく、悪性の部分と良性の部分に分けて人の手でマーキング、線引き、ROI（Region of interest）で囲むなどの作業を行う。機械学習用に細かいパッチにした後では良性、悪性の鑑別が困難となることも少なくなく、作業が慎重に進められるため、1 枚の WSI につき平均的に約 15〜20 分程度かかる。20 万枚にアノテーションを付与するのに、1 枚 20 分として 66,667 時間、1 人の病理医が 1 日 8 時間アノテーション付与作業に専念したと仮定した場合、仮に 100 人の病理医で取り組んだとしても、すべての画像にアノテーションを付与するのに約 2 か月半の時間を要する。教師なしデータを使った AI 病理診断支援プログラムの開発研究に関する論文ベースの報告もなされているが、日常診療で使用する病理診断支援プログラムという観点からは実用できる段階には到達していない。このアノテーション付与という人の手による律速段階の作業効率化に対する技術革新も WSI データを活用する上での大きな課題であろう。

想定される利用方法

病理医が期待している病理診断支援のための SaMD はいくつかある。現在、病理学会として診療報酬改定で保険収載を要望しているのが、病理診断の「ダブルチェック」を行う SaMD である。日本では慢性的に病理医が不足しており、病理医が不在の医療機関も多数ある中、病理医が常勤で勤務する医療機関でもその約 45%が、1 人しか病理医が勤務していない「1 人病理医療機関」となっている。この 1 人病理医療機関では、疾患の最終診断である病理診断がダブルチェックされずに報告されており、その精度管理の体制や、1 人病理医の精神的負担の軽減等が病理学会の長年の課題となっている。例えば病理診断の多くを占めている消化器内視鏡検体の病理診断のダブルチェックを行う SaMD は、精度管理と精神的負担軽減の両面で 1 人病理医の支援となることが予想され、この SaMD 開発のための WSI のアノテーション付きデータベースの活用が期待される[41]。またまだ研究開発ベースであるが、HE 染色標本から遺伝子変異を予測する SaMD 開発のため、遺伝子変異と結びついたデータベースを構築している団体や、希少がん WSI データベースを活用して、HE 染色標本の類似画像から診断名を想定し、確定に必要な免疫染色や遺伝子変異検索を提案するという SaMD などの開発のためのデータベースの構築にも取り組みがなされている[43]。

病理デジタル動画データベースにおける課題

AMED の研究支援で収集した WSI に電子的にメタデータ等が含まれていないかなどの確認を「個人情報及び匿名加工情報扱い委員会」で行ったが、かなりの時間を要

した。また先述のようにWSIには、フォーマット形式が複数あり、その各フォーマットに対応した各社個別のビューワソフト等の関係等で、保管されている WSI のうち 12.2 万枚の病理組織画像に関してのみが公開可能な状況となっている。今後は、統一規格のフォーマットや、汎用ビューワソフトの開発が待たれるところである。

７－３．心電図データベースの概要

データベース構築の目的

心電図データベース構築の主な目的は、自動診断の精度を上げ、診断の補助として役立てることや、発作性の疾患を対象として非発作時の心電図より有病の有無を予測することが挙げられる。後者の場合、アノテーションデータは心電図所見の中にはなく、別に用意されることが特徴となる。

収集データの種類・規格

心電図データは、メーカーにより異なるフォーマットで記録されているが、学習データとして使用する場合は時系列のテキストデータに変換して使用することが多い。標準 12 誘導心電図であれば 12 列×時間点により構成される行列データとなる。しかし、心電図波形を画像としてとらえる手法もあり、その場合は個別の誘導または 12 誘導を総体として画像データとして収集することもある。

現状の収集データ規模

発作性不整脈疾患の有病予測の一例として、発作性心房細動の有病予測のため、東京医科歯科大学・自治医科大学を中心として 7 施設 2,700 例のデータ収集を行った。収集時の心電図は正常洞調律の標準 12 誘導心電図であり、アノテーションデータは心電図のテキストデータとは別に、発作時心電図の記録の有無として用意され紐付けられている。ここで作成したデータの特性は、心電計は同一機種で統一し、多施設（大学病院 2、循環器センター病院 5）かつ異なる環境（病院検査室、病室）で収集した。また、抗不整脈薬の服用など、結果に影響を与える可能性があるものを明確に除外しているのも特徴である。ここで収集した心電図データ利用の注意点としては、発作性心房細動あり、のアノテーションには実際の心房細動発作の記録があり、診断が確実であるのに対し、発作性心房細動なし、の群では過去生涯にさかのぼる心電図の全記録がないため、無症状の発作性心房細動が見逃されている可能性があることである。この問題点は、発作性疾患の有病予測・発症予測という形でデータを利用する場合のコントロールデータの限界点として常に生じるものであり、注意が必要となる。

紐付けられる医療情報の種類

心電図データに紐づく情報として患者背景、術者情報（心電図電極が正確か担保する意味で、臨床検査技師または医師が記録したものであることが望ましい）、機器情報（心

電計による特性の評価)、環境情報（検査室の外部ノイズ、室温、湿度)、患者情報（前述の発作性不整脈疾患の有無に関するアノテーションデータ)、心電図情報に反映される可能性のある併存疾患・既往歴の有無、抗不整脈薬の服用、血液生化学検査による電解質データなどがある。また、心電図の自動診断に用いる場合は、専門医による心電図所見のデータが必要となる。

異なる個人、地域、医療機関、国、時期におけるデータ収集の必要性

心電図の検査手技自体は標準化されているが、異なる環境での計測による外部ノイズの混入による影響を考慮し、多施設・異なる環境でのデータ収集が必要である。また、アノテーションデータの観点からは、疾患分布の地域差などを考慮したデータ収集が必要と思われる。

人の手技に対する依存度の有無

心電図検査手技は標準化されているが、熟練していない医療者による心電図データは、不正確な電極位置での記録によりデータの質が変化することがある。データベース作成の観点からは、臨床検査技師または循環期内科医師など、測定者を限定することが望ましい。

撮影・記録に関わるデバイスの多様性

医療機器承認を得ている心電計のうち、12 誘導心電計に関しては、データに一定の質が担保されていると考えられるが、計測条件によりサンプリングレートや量子化ビット数の違いが生じることがある。また、ホルター心電計については、装着位置により心電図信号の質に大きな影響がでることがあり、特に近年利用範囲が広がっている小型 1 チャンネル長期間装着型心電計において、その傾向が強い。

使用デバイスの多様性

心電図計測における電極は、導電性ゲルを用いたシール型電極、クリップ型電極・吸着型電極などの種類があるが、シャツ型電極など導電繊維を用いた電極などの開発が進んでいる。電極の特性により心電図信号の質に影響を与える可能性が将来的には考えられる。

主なアノテーション手法・教師データの作成方法

心電図データ収集時は、データ取得時点でのアノテーションデータが付与されるが、その後に新たな不整脈発作を生じた場合に、該当疾患に関するアノテーションデータが新たに付与されることになる。このとき、該当心電図を、不整脈発作群として分類するか、同一被験者の反復測定心電図データのなかで、疾患発症前の記録としてアノテーションデータを付与するか、という問題が生じる。この点に関する明確な基準はなく、解

析手法によってアノテーションが変わることも十分に考えられる。このため、アノテーションデータとして不整脈発作の有無を与える場合には、心電図記録日時と不整脈発作の出現日時を含めてデータベース化する必要がある。

想定される利用方法

心電図データベースを用いたモデル構築の利用方法は大きく 2 点あり、①記録時に生じている心電図波形変化を自動診断する、②記録時には生じていない発作性不整脈の発症を予測する、という活用が考えられる。①の利用法においてはアノテーションが心電図記録の中に含有され、確定したアノテーションデータが得られる一方、②においてはアノテーションデータは心電図記録とは別にあるという特徴がある。

心電図データベースにおける課題

心電図データの規格は世界的には統一されておらず、通常テキストデータへの変換を通じて収集されるが、時間解像度（サンプリングレート)、ダイナミックレンジ、量子化ビット数は機種及び記録条件により異なる。記録時間についても、標準 12 誘導であれば 10 秒間が一般的であるが、統一はされていない。

心電図データの収集は比較的容易であり、同一被験者から複数回のデータ取得が可能である。しかし、同一被験者からのデータは必ずしも同一データの再現を見ている訳ではない。長期的な時系列変化として、加齢や背景疾患の進行により心電図情報が変化することが考えられるほか、短期的な変化としても、日内変動や日差変動といったレベルでの変動が心電図に反映される。さらに、心拍に応じて波形が繰り返し記録されるという心電図の特性から、一拍ごとの心電図波形の微小な変化に特徴量が含まれていることも想定される。これらを時間変動と呼ぶが、時間変動は静的なデータとしての心電図とは異なる情報をもつ。以上より、異なる時相で得たデータは、完全な独立データではないものの異なる情報をもったデータとしての扱いが必要になるほか、同じ時相で得たデータであっても短期的な時間変動を含有したデータとしての扱いが必要になる。

前述のように、発作性不整脈疾患に関する正確なアノテーションデータを付与するには、生涯にわたる心電図連続記録を基にした診断が必要となる。

７－４．消化器内視鏡画像データベースの概要

データベース構築の目的

データベースを構築する目的は、消化管内視鏡検は直接診断のために記録する。治療目的の内視鏡に関しては、今後利用目的が出てくる可能性はあるものの、外科的手術の動画のように技術の客観性の評価のような目的に運用されることは今のところ進んでいない。専門医に伍するような精確な診断ができるようになることで、消化器内視鏡診断が一般化、強化されることが大きな目的であるといえる。ただし、ここでいう診断とは単純に疾病名、診断名を判断するあるいは発見するという目的にとどまらず、病期な

どの判定や、偶発症予測など、複雑な判定を含む方向に進んでいる。

収集データの種類・規格

消化器内視鏡領域において収集すべき対象は、静止画と動画の双方ということになる。静止画はかなり多くの施設ですでに集積されており、この十分なリソースをどのように用いるかが大きな課題であり、消化器内視鏡領域の特異性であるといえる。規格としては現状展開されている消化器内視鏡画像ファイリングシステムにおいて、JPEG やビットマップなど汎用的な規格で補完されており、運用の展開は容易であるといえる。しかしながら後述するが画像のみでは収集データのリソースとしては不十分である。内視鏡画像には大量の画像それぞれに意味づけが可能である。撮影されたものが胃癌なのかどうかという単純な初期的命題にとどまらず、消化器内視鏡診療において診断され得る非常に多くの病変の診断情報をタグ付けとして付加し集積することこそが重要で、診断名のみならず、疾患背景などを標準化した JED（Japan Endoscopy Database）プロジェクトで規定した標準化用語をタグとして付加することで研究や運用の幅が出てくることを強調したい。

一方動画に関しては確実なフォーマットは存在しない。しかしながら消化器内視鏡学会においては、動画から静止画を切り出し、切り出し静止画像に対して、簡便なタグ付けや、膨大な生成静止画像に対してアノテーションが可能なるツールを開発しており、標準化を義務化することよりも、研究者開発者にとって利便性の高いツールを流布することで、緩やかな標準化を目指すべきであると考えている。

現状の収集データ規模

AMED の支援やその他の開発目的に集積された画像は、各分野により異なるものの、多施設での検討において、胃癌症例の場合は 3 万画像に加えて正常画像 13 万画像、炎症性腸疾患においては 1 万画像、胃の撮影部位の認識においては当初のアルゴリズム作成のために 2 万画像に至る。十二指腸乳頭部画像は 54 万画像と膨大な画像を複数領域にわたって集積した。一方で消化器内視鏡分野においては商用化がなされているものもあり、それらは日本消化器内視鏡学会として集積した画像とは別に集積されており、全貌をここで述べるのは困難である。消化器内視鏡領域においては、すでに企業が商用化に至った製品、商用化を目指している製品があるため、一意にすべての画像が集積されている訳ではなく、企業の論理や研究者が扱える範疇にとどまって運用されているものもあり、全般的な統合データベースを作成するのは非常に困難であるといえる。動画像からの静止画切り出しに関しては大腸内視鏡検査において 10 万画像を生成して研究に供した。

紐づけられる医療情報の種類

静止画、動画に関わらず、AI 学習に用いる形態としては静止画を基準として、紐づけ

データを考慮すればよいと思われる。先に記載したように日本消化器内視鏡学会として標準化したJEDデータ用語に準じて、詳細な情報がタグ付けされることが重要である。また消化器内視鏡においては、撮影された画像がどの臓器のどの部位を撮影したものかという情報も重要なものとなる。特に部位の自動認識をする際にはLocation情報は重要な点であり、開発、研究の目的次第で項目を付加していくことも重要な観点であり、前述した動画像から静止画を切り出すツールでは付加可能なタグ情報の追加にも対応したものとなっている。

JEDデータとしては大きく診断情報として、質的診断（病名）のほか、部位、大きさ、存在所見（内視鏡的な所見情報）があり、これに加えて患者背景情報として、家族歴、既往歴、嗜好に関しても標準化した用語を提供しており、JEDフォーマットに準拠して保管されることで整理された情報が得られる。さらに消化器内視鏡特有の事象として、内視鏡（スコープ）の機種や撮影状況なども標準化されている。

異なる個人、地域、医療機関、国、時期におけるデータ収集の必要性

疾病には地域的に罹患数が多いものがあり得る。これは国内外を問わない。その上異時性のある情報は非常に多くの意味をもち、同一患者においても、経時的に変化してゆく所見は十分にあり得るもので、同一個人であっても異時性データは違うものとして扱う必要がある。さらに正常という概念は研究、開発において、非常に重要な概念であり、病名がないものという考え方は取らず、正常は正常としてとらえた上で集積する必要があることは強調しておきたい。

もう一点、消化器内視鏡は1検査で最低でも40画像程度の静止画を保管している。希少疾患においては1検査で100枚以上の画像が集積される。しかしながら、一つの検査というアクションの中で病変が存在したとしても病変以外の画像が含まれた形で集積されるのが大きな特徴である。現状では検査単位の情報のタグ付けにとどまっている状況であるが、動画から静止画を切り出すアプリケーションのごとく、学習すべき全画像に個別のタグ情報が必要になるのが大きな特徴である。この点は病変が撮像されたものだけを特異的に用いる他領域と大きく異なる点であり、一方で大量の画像をタグ付けが詳細になされれば簡便に得られることとなり、この特徴をどう活かすかは今後の課題であるといえる。

人の手技に対する依存度の有無

放射線機器や顕微鏡などのように、客観性の高い画像に比して、消化器内視鏡をはじめとした各種内視鏡画像や超音波画像においては、撮影者の技量や経験にかなり大きく依存するため、何らかの層別化は重要である。外科手術動画のように、経験年数などが妥当なファクターではあるものの、患者の個人情報に加えて医師側の個人情報をどのように扱うかも大きな命題であり、今後の課題であるといえる。

撮影・記録に関わるデバイスの多様性

内視鏡（スコープ）の機種名や光デジタル法と呼ばれる様々な特殊光観察の条件情報は重要であるといえる。しかしながら特殊光観察は 1 検査の中で頻繁に条件が変わっていくという特徴もある。すなわちこの点においても、検査単位ではなく、画像 1 枚 1 枚の単位で条件がタグ付け情報として付加されるのが理想であり、その方策の策定も徐々にではあるが進んでいる。

使用デバイスの多様性

検査内視鏡だけでもいろいろなデバイスが運用される。われわれの検討ではデバイスそのものを認識させる取り組みも行っている。

主なアノテーション手法・教師データの作成方法

AI 研究に付与する画像処理、学習の方法論により異なるとは思われるが、1 枚の画像全体に病変のみが撮影されるようなことはほぼなく、1 枚の画像の中の一部に病変が映っているのがほとんどである。そのため病変認識においては、病変部を正確にトレースするアノテーションが重要となる。

想定される利用方法

本データベースの想定される利用方法としては、①Detection（病変指摘）、②Different Diagnosis（鑑別診断）、③Deviation（逸脱監視）（規定されたルールに則った検査が完遂されているかのチェック）、④Staging（病期診断、治療効果判定も含む）、⑤Prediction（手技の難易度や偶発症発生率）などの予測が挙げられる。

消化器内視鏡画像データベースにおける課題

課題として最も大きいのは倫理的問題である。研究目的に集積した画像は、当初掲げられた目的以外への利用は禁じられており、本来的になんにでも利用できるようなデータベースとはなっていない。個別同意を得た運用自由度の高いデータベースを構築していく必要がある。

７－５．データベースにおける共通の課題・問題点

データの容量が大きいこと

全てのデータベースに共通して言える課題であるが、膨大なストレージ容量を要する。例えば、手術動画 FHD1 本 3 時間の場合、18GB 程度の容量である。4000 本を保存する場合、72TB が必要になる。動画の規格を統一する前のデータのバックアップや個人情報に配慮した加工を行うとなると、倍以上の保存容量を要する。保存容量が大きいことのデメリットは、学習やテストを行う際に、時間がかかること、クラウド管理をする場合、多額の費用がかかることである。

アノテーションデータセット作成の煩雑さ

医師の診断や判断が必要であり専門性が高く、アノテーションはかなり専門的かつ複雑である。例えば、消化器内視鏡学会では消化器内視鏡における研究に資するような、アノテーションツールを作成したものの、膨大な時間を軽減する程度である。

データベースに関連する様々な費用の捻出

データベースにおいては、構築のみならず、運営、管理などの費用の捻出が大きな課題であるといえる。日本国内のデータベースの多くは、AMED などの研究費を使用し、収集しているが研究終了後にその維持運営費用が途絶えてしまう。その意味において今後、研究費に依存せず事業化し持続的に収集していくことが大きな課題の 1 つである。

病理デジタル画像のデータベースでは当初、企業が運用するセキュアなクラウドでの保管を計画したが、年間の見積もりが数千万円であったため、現在はサードパーティーの年間数百万円程度のクラウドストレージに保管し、運用に関しては病理学会会員が自ら行っている状況である。画像の活用に関しては病理学会会員向けには、生涯学習用の人利用や会員が行う基本的な学術研究目的の AI 開発用の画像解析に関しては、日本病理学会研究委員会で研究計画書等を確認の上、無償での画像活用を認めている。一方、クラウドの運用費用捻出のため、今後は、企業が研究費を出資するようなプロジェクトや教育目的でも企業等が使用する場合、あるいは企業自らが AI の開発を行う場合には、アノテーションのない WSI、1 枚当たり 500 円などで後提供すること等で、クラウドの運用費捻出を計画している。手術動画データベースにおいては、AMED 事業終了後に独自の運営会社を設立データベースの維持、運用を継続している。

８. 深層学習等の機械学習を用いた SaMD の開発のためのデータ（学習データ、検証データ、テストデータ）に関する考察

SaMD の開発のためのデータには、学習（training）、検証（validation）、テスト（test）の 3 種類のデータがある。学習は例えば深層モデルの重みパラメータの学習に用いるデータであり、検証は深層モデルの構造や学習回数などのハイパーパラメータを決定するために用いられる。広い意味ではこれら 2 つが学習用データである。一方、テストデータは、学習済みの SaMD を市販後に real world データに適用した際の性能を推定するために用いられる。

これら 3 種類のデータは主にモデルの開発の段階で、交差検定（cross validation）などのプロセスにおいて使用される。データに求められる条件を考える場合にこれら 3 つのデータのうちどのデータについての議論を行っているのかを明確に意識しなければならない。ここでテストデータについてはさらに、SaMD の医療機器としての承認審査・認証のプロセスで利用され、当該 SaMD が想定する医療応用の場面において十分な性能を有しているかを評価するためのデータが存在する。以下の考察における条件や仮定は、開発で使用されるテストデータより後者の承認審査・認証のテストデータの方に対して、より厳密に適用されることに注意されたい。

機械学習の分野では、学習用データ（学習と検証の両方を含む）がテストデータに対して独立同一分布であるという仮定、すなわち、適用を想定している real world のデータから独立してサンプリングされている同一の分布であるという仮定を設定することが多い。以下ではこの学習用データとテストデータの間の分布の独立性と同一性の仮定について考察する。

独立性

開発時に使用する学習・検証用データと開発した深層学習モデルの性能評価に用いるテストデータが独立であることは、テストの妥当性を維持するために必須の条件であることは言うまでもない。このように学習用データとテストデータの独立性の仮定については、完全に同一のデータを学習とテストの両方で利用することを禁止することはもちろん、テストデータに類似した画像を学習データに意図的に加えるなど、独立性に疑義が生じる行為については厳しく禁止する必要がある。しかし、その考え方を安易に延長して、同一被検者のデータを学習とテストの両方に分けることを一様に禁止すべきか否かについては、十分な議論が必要である。

このことは前述のデータベース開発の事例においても議論されている重要な論点である。独立性を強調するあまりに同一被検者のデータの利用について過度に強い制限を与えると、貴重な医療データが有効に利用できなくなる可能性がある。例えば、同一被検者のデータでも、疾患の種類や発生部位、データの取得時期が離れており、データの特徴が大きく異なる場合には、同一被検者であったとしてもデータ間の独立性が例外的

に担保できる可能性もある。同一被検者のデータ間の独立性については、疾患の種類、部位、撮影時期などの観点から慎重に判断し、条件付きで適宜緩和する選択肢を残す必要がある。なお、条件を緩和した場合には、学習データとテストデータの間の独立性に疑義が生じないように、その妥当性について科学的・客観的に説明することが必須となる。開発が終了したモデルの性能を正しく推定するためには、開発時に使用した学習・検証用のデータとのテストデータの独立性が維持されていることが重要となる。

学習用データとテストデータの間の分布の同一性

学習用データ（学習と検証の両方を含む）がテストデータに対して独立同一分布であるという仮定に関する議論は上記３．バイアスの議論と関連する。また上記７．で紹介されている様々な分野でのデータベース作成の取り組みは、応用分野を定めれば妥当性をもって同一性を説明できるデータを提供できるものとなっている。今後新しい分野での機械学習用データと提供するデータベース開発においては、これらの議論の視点からのデータベースの仕様決定を進めることが望ましい。

学習完了後の深層学習モデルの性能を評価するために用いられるテストデータに関しては、そのテストデータが SaMD の適用を想定している real world の分布から学習時に使用されたデータとは独立してサンプリングされているという仮定が設定されていることが求められる。一方、学習段階においては、学習データが real world の同一分布からサンプリングされているという同一性の仮定を厳密に適用すべきかに関しては異なる議論が必要である。最近、学習段階においては、この同一性の条件を満たさない学習データを併せて使用することが機械学習による性能向上につながる可能性があることが報告されている。例えば、大量の自然画像などで学習させたモデルを医用画像に適用した例は多数報告されている[44, 45]。加えて、基盤モデル（foundation model）[46]と呼ばれる桁違いの大量のデータ（医用画像以外の自然画像やテキスト情報など）を使って学習したモデルも登場している。例としては、Segment Anything Model (SAM)のように 10^7 枚以上の画像や 10^9 以上のセグメンテーションマスクを用いて設計され、さまざまな種類の画像に適用可能な高い汎用性を持つモデル[47]である。例えば画像分類の問題で医用画像ではない大量の自然画像に対する学習を行った上で、医用画像データによる学習を行わせることで優れた性能が達成されるといった、学習用データがテストデータと同一分布であるという仮定が成立しない場合も報告されるようになった。深層学習を利用しない SaMD の多くは、実際に適用を想定する応用分野における real world の同一分布から独立してサンプリングされた学習用データを用いて設計されることがほとんどであった。これは前述の様々な医療分野でのデータベースを活用した医療用深層学習モデルの開発においても、当たり前の条件であると考えられてきた。しかし、最近は事情が変わりつつある。今後は、非医用画像で学習したモデルのように、学習用データとテストデータの間に同一分布の仮定が成り立たないケースが増えると予想されるため、今後の SaMD の審査においては、開発時に使用される学習用データ・検証用データ

が想定する応用分野の real world データと同一分布を持つという仮定に強くこだわることは適切ではないと考えられる。

また、転移学習（Transfer Learning）[48]やドメイン汎化（Domain Generalization）[49]などのように、そもそも学習用とは異なる分布に対応するための手法も登場するなど、深層学習などの人工知能技術は高い柔軟性を備えるように進化していることからも、学習用データと real world データの間の分布が同一であるか否か、あるいは学習用データとテストデータの間の分布が同一であるか否かについては過度にこだわる必要はない。より重要な点は、当該 SaMD が所望の性能を有しているかどうかを、real world データに適用した際の性能を用意したテストデータを使って適切に推定できるか否かである。そのためには、テストデータは SaMD の適用を想定している real world のデータから学習時に使用されたデータ（検証用データも含む）とは独立してサンプリングされているという仮定や、４．で指摘したテストデータの再利用においてテストデータの情報が開発側に漏れないためのさまざまな工夫がより重要であることを再度強調しておく。

深層モデルの仕様不足（underspecification）

最後に、最近指摘されるようになってきた、深層モデルの仕様不足（underspecification）[50]の問題について触れておく。

本項の最初に述べたように、テストデータは、学習済みの SaMD を市販後の real world データに適用した際の性能を推定するために用いられ、性能を正しく推定するためには、このテストデータと real world データの間に分布の同一性が成り立つことが重要となる。

しかし、実際には市販後の real world データに対する性能が異なるにもかかわらず、設計側の手元にあるテストデータではその性能の区別がつかない複数のモデルが存在する。そのため、開発時には同じ性能を示した複数のモデルが実際の real world データに適用した場合に異なる性能を示すケースがあることが報告されている。これは仕様不足と呼ばれる問題として位置づけられている。この問題を抱えた SaMD を市販後のデータに適用した場合、承認時と比べて性能が大幅に低下するおそれがある。仕様不足の評価法としてテストデータに意図的に摂動を与え、データの近傍にばらつきを与えてテストを行う Stress Test と呼ばれる方法などが提案されているが[51]、その対応策については未解決の課題として現在も研究が進められており、評価が確定していない点も残されている。

今後もこの仕様不足の研究の動向については引き続き注意が必要である。

９．まとめ

本報告では、機械学習を応用した SaMD を開発し、医療機器ソフトウェアとして社会実装する上での課題を議論し、主にデータに関して考慮すべき項目をまとめた。

「バイアス」という観点で様々な考察が必要であることが示されている。すなわち学習、検証や評価に用いられるデータが、その SaMD の意図する使用において対象とする患者集団の統計的性質と同等であることが求められる。また、利用者が意識せず持つバイアスや、データを選択する場面、アノテーションを付す場面でヒトが潜在的に有してしまうバイアスといった、これまであまり着目されていない点も指摘した。これらの考え方は、各国規制当局が推奨する“Good Machine Learning Practice”という概念と共通するものである。

一方、市販後に実世界データを集積し学習を継続することで性能を変化させるように設計することも可能という点が、機械学習を用いた SaMD の特徴の一つである。このような製品をどのように評価するのかといった問題に対して、米国 FDA は Pre-Cert を試行しているが、現状では多くの課題が指摘されたことを述べた。このことは市販後学習の手法を明確化したとしても、市販前に市販後の性能を科学的に推定することの難しさが現段階では存在することを示している。また、学習を行う場合に実世界での性能評価結果を基に、再学習に用いるデータを選択する場合にも、バイアスが発生する可能性があり注意が必要である。このような問題を回避する手法開発研究も含め、この問題に対する現段階の科学的知見を本報告では紹介した。

医療分野における機械学習応用の難しさの一つとして、一定の品質を有する学習、検証、評価データを大量に収集することの困難性が指摘されている。公開されている他の用途で収集された大量のデータを用いて学習することも、この分野ではしばしば行われる。これらデータが固有に有する、人間には認識できないレベルのバイアスが学習データとして開発された AI モデルの性能に影響し、開発段階で実現された性能と実世界データに対する性能との間の乖離につながる可能性があることを、文献を引用して指摘した。また、学習データを得る新たな方法として、数値解析を用いたシミュレーションデータの活用を本報告で議論し関連する課題を指摘した。数値計算モデルの妥当性の検証など、過去に科学委員会が公表した「コンピューターシミュレーションを活用した医療機器ソフトウエアの審査の考え方に関する専門部会報告書」[3]に示された考え方が活用できる。昨今の AI による画像生成技術の進歩により、学習データの少なさを補うために画像生成技術を応用することも今後研究されるものと推測される。いずれにせよ学習/検証/評価において使用するデータの（統計的な）性質がその SaMD の意図する使用において対象とする患者集団の統計的性質と同等であることが求められるであろう。

また、現在科学論文として公表されている本分野関連研究では、多くは非無作為化試験が前向きではなく、バイアスのリスクがあることを文献研究により指摘した。一定の品質を有する評価データを大量に収集することが難しい医療分野において、限られたデータを用いて科学的に適切な評価を行うかについては、従来の医療技術評価をもとに本

報告で指摘した多くの留意点を参照しながら個別に合理的な考察を行うことが求められるであろう。

さらに、機械学習に応用することを意図した医療データのデータベース構築の現状を複数の分野で調査し、本専門部会での議論を基に共通課題を抽出した。指摘された留意事項は、今後医療データベースの強化、新たなデータベース構築を行う際のデータベース設計・運営にあたって参考となる内容である。

データには機械学習モデルの学習に使用される学習データ、いわゆる教師データ、開発プロセスの中で学習過程により得られたモデルの性能を検証し、さらに開発プロセスの中でモデルの改良に使う検証データ、そして開発プロセスの中でのテストデータの 3 種がある。そして完成された機械学習モデルを医療機器として承認や認証を得るために、その性能を評価するためのテストデータがある。この点は第 8 章において関連する内容を「深層学習等の機械学習を用いた SaMD の開発のためのデータ（学習データ、検証データ、テストデータ）に関する考察」として論点をまとめた。ともするとこれら 3 種類のデータについて一体で考え、同一の品質を求めてしまいがちであるが、バイアスを適切に管理すれば学習データにシミュレーションデータを使うことも可能であると考えることや、想定する応用分野で得られるデータとの同質性が保たれていないデータによる学習（例えば医用画像処理における大量の自然画像データを用いた学習など）が結果として優れた性能を得るために有効であるという知見の存在など、何の目的で使われるデータであるのかを意識しながらその要求品質を定めることが必要である。しかし、当然のことながら深層学習等の機械学習を応用した SaMD の性能評価の目的で使用されるテストデータとして、シミュレーションデータを評価データに使うことは、シミュレーションデータがテストで評価すべき目的に即して現実の現象を適切に再現できるという妥当性を有することが科学的に示されない限り、現段階では受け入れられない。現段階で本報告ではテストデータに関しては、real world と同一分布からテストデータが学習用データ（開発時に使用された検証用データも含む）と独立してサンプリングされているという仮定を設定すべきであるという立場をとっている。高い品質の医療データを大量に取得することが難しいという医療データの特性を鑑みると、データに求められる要件に関して今後さらに考察を深める必要がある。

作成したデータベースに記録されたデータを、開発した SaMD の承認申請データとして使用することを想定する場合は、以下の点に注意しなければならない。

- 個人情報保護法、臨床研究法等の関連法規に十分配慮し、信頼性調査を含む審査のプロセスにおいて、必要になる資料が提出できるよう準備しておく必要がある。これらの当該関連法規に抵触することにより、必要なデータにアクセスできない結果、審査に必要となる十分な資料の提出や考察ができない場合には審査が困難となる可能性があり、この点十分注意しなければならない。
- 収集された医用画像などの情報や、それに紐づく臨床情報を利用することに関しては、承認申請を含む製品開発（商用）に当該データが利用されるというデータ

の使用目的について患者の同意が得られた上で、適切な方法でデータの匿名化が行われている必要があることに留意すること。

医療分野への機械学習を用いた AI の導入に対する期待が高まっているが、本報告で示された留意点を参考にしながら、その適切な開発プロセスの設計と、科学的かつ合理的な性能評価を行うことが、安全で有効な深層学習を用いた AI 医療機器の社会実装につながるものと考えられる。

【参考文献】

[1]. 科学委員会. AI を活用した医療診断システム・医療機器等に関する課題と提言 2017. 2017（閲覧 2023/7/26）https://www.pmda.go.jp/files/000224080.pdf

[2]. プログラムの医療機器該当性に関するガイドラインの一部改正について（令和 5 年 3 月 31 日 薬生機審発 0331 第 1 号、薬生監麻発 0331 第 4 号）. 2023（閲覧 2023/7/26）https://www.mhlw.go.jp/content/11120000/001082227.pdf

[3]. 科学委員会. コンピューターシミュレーションを活用した医療機器ソフトウエアの審査の考え方に関する専門部会報告書. 2021（閲覧 2023/7/26）https://www.pmda.go.jp/files/000240657.pdf

[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（閲覧 2023/7/26）https://www.fda.gov/media/122535/download

[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（閲覧 2023/7/26）https://www.fda.gov/media/166704/download

[6]. U.S. Food and Drug Administration. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. 2021（閲覧 2023/7/26）https://www.fda.gov/media/145022/download

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

[8]. U.S. Food and Drug Administration. The Software Precertification (Pre-Cert) Pilot Program: Tailored Total Product Lifecycle Approaches and Key Findings. 2022（閲覧 2023/7/26）https://www.fda.gov/media/161815/download

[9]. 国立医薬品食品衛生研究所医療機器部. 令和 3 年度 AI・モバイル用アプリケーション等最先端医療機器調査等事業報告書. 2022（閲覧 2023/7/26）https://dmd.nihs.go.jp/samd/R3_report.pdf

[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（閲覧 2023/7/26）https://www.fda.gov/media/153486/download

[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 (閲覧 2023/7/26)

[12]. 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 薬生機審発 0523 第 2 号）. 2019（閲覧 2023/7/26）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

[13]. 独立行政法人医薬品医療機器総合機構. 医療機器プログラム（SaMD）の審査ポイント. （閲覧 2023/7/26）https://www.pmda.go.jp/review-services/drug-reviews/about-reviews/devices/0047.html

[14]. プログラム医療機器の特性を踏まえた適切かつ迅速の承認及び開発のためのガイダンスの公表について（令和 5 年 5 月 29 日 事務連絡）. 2023（閲覧 2023/7/26）https://www.mhlw.go.jp/hourei/doc/tsuchi/T230530I0080.pdf

[15]. IMDRF/AIMD WG/N67. Machine Learning-enabled Medical Devices: Key terms and Definitions. 2022（閲覧 2023/7/26）https://www.imdrf.org/sites/default/files/2022-05/IMDRF%20AIMD%20WG%20Final%20Document%20N67.pdf

[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（閲覧 2023/7/26）https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/

[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]. 医療機器の変更計画の確認申請の取扱いについて（令和 2 年 8 月 31 日 薬生機審発 0831 第 14 号）. 2020（閲覧 2023/7/26）https://www.pmda.go.jp/files/000236900.pdf

[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]. 吉村功.検証的比較臨床試験の計画において考慮すべきこと-ICH 統計ガイドラインの理解のために-. 統計数理.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]. 佐久間淳.『データ解析におけるプライバシー保護』講談社.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 CM 著, 元田浩ほか監訳. パターン認識と機械学習　上. シュプリンガー・ジャパン. 2007;p.11.

[31]. 医薬・生活衛生局医療機器審査管理課. 審議結果報告書. 2022（閲覧 2023/7/26）https://www.pmda.go.jp/medical_devices/2022/M20220516002/112714000_30400BZX00101_A100_4.pdf

[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]. 佐々木毅. 日本病理学会 JP-AID と病理診断人工知能開発. 病理と臨床 2017;35(11)

1058-1061.
[39]. 宇於崎宏ほか. 第 7 回 日本病理学会でのデジタル画像収集基盤整備. 病理と臨床 2018;36(10):1017-1021.
[40]. 一般社団法人 日本病理学会 日本デジタルパソロジー研究会 デジタルパソロジー技術基準検討会. 病理診断のためのデジタルパソロジーシステム技術基準 第 3 版. 2018（閲覧 2023/7/26）https://pathology.or.jp/news/pdf/kijjun-181222.pdf
[41]. 佐々木毅. 病理診断支援 AI アルゴリズムの開発：日本病理学会の取り組み. 医療機器学. 2019;89(6):526-532. (doi:10.4286/jjmi.89.526)
[42]. 佐々木毅.病理診断領域における AI プログラムの課題と展望. Modern Media. 2022;68(3):74-80.
[43]. 佐々木毅. AI による病理画像診断. 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. （閲覧 2023/7/26）(https://jmlr.org/papers/volume23/20-1335/20-1335.pdf)
[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)

AIを活用したプログラム医療機器に関する専門部会　委員名簿

○	伊藤 雅昭（いとう まさあき）	国立がん研究センター東病院　副院長・大腸外科長・医療機器開発推進部門長
◎	佐久間 一郎（さくま いちろう）	東京大学 大学院工学系研究科附属医療福祉工学開発評価研究センター 教授
	佐々木 毅（ささき たけし）	東京大学 大学院医学系研究科 次世代病理情報連携学講座　特任教授
	笹野 哲郎（ささの てつお）	東京医科歯科大学 大学院医歯学総合研究科 循環制御内科学分野　教授
	澤 智博（さわ ともひろ）	帝京大学 医療情報システム研究センター　教授
	清水 昭伸（しみず あきのぶ）	東京農工大学 大学院工学研究院　教授
	陣崎 雅弘（じんざき まさひろ）	慶應義塾大学 医学部放射線科学教室（診断） 教授・診療部長
	武田 理宏（たけだ としひろ）	大阪大学 大学院医学系研究科 情報統合医学講座 医療情報学　教授
	田中 聖人（たなか きよひと）	京都第二赤十字病院 内科部長・医療情報室長
	鎮西 清行（ちんぜい きよゆき）	産業技術総合研究所 健康医工学研究部門 首席研究員
	殿村 桂司（とのむら けいじ）	長島・大野・常松法律事務所 弁護士
○	中岡 竜介（なかおか りゅうすけ）	国立医薬品食品衛生研究所 医療機器部 埋植医療機器評価室長
	中田 典生（なかた のりお）	東京慈恵会医科大学　人工知能医学研究部　教授
	中田 はる佳（なかだ はるか）	国立がん研究センター 研究支援センター 生命倫理部 COI管理室長
	村垣 善浩（むらがき よしひろ）	神戸大学　未来医工学研究開発センター　センター長 / 大学院医学研究科・医学部　教授
	森 健策（もり けんさく）	名古屋大学　大学院情報学研究科　知能システム学専攻　システム知能情報学　教授/ 名古屋大学　情報基盤センター長
	横井 英人（よこい ひでと）	香川大学 医学部附属病院 医療情報部　教授

◎部会長、○副部会長
（五十音順）