第４回ＡＩを活用したプログラム医療機器に関する専門部会

日時　　令和５年２月２０日(月)
　　１４：００～
開催形式　Ｗｅｂ会議

＜開会＞

○事務局（渕岡先端技術評価業務調整役）　それでは、第 4 回 AI を活用したプログラム医療機器に関する専門部会を開催いたします。本日はお忙しい中、お集まりいただきありがとうございます。

＜委員出席状況報告及び配付資料確認等＞

○事務局（渕岡先端技術評価業務調整役）　はじめに委員の出席状況について報告いたします。当委員会の 17 名の委員のうち、現在 13 名の委員に御出席いただいており、全委員の過半数に達しておりますので、専門部会規程第 7 条に基づき、本委員会の成立を御報告いたします。

次に、配付資料の確認をさせていただきます。画面に表示させていただいておりますが、議事次第・資料目録、その他に資料取扱区分表があります。資料としては、資料 1～5 及び参考資料があります。資料に不足等がございましたら事務局までご連絡をお願いいたします。

次に、資料取扱区分表を御覧ください。資料は内容に応じて、取扱いとして「厳重管理」「取扱注意」「その他」に分類し、それらに応じた対応を取ることとしております。本日の配付資料のうち、資料 1 と参考資料 1 は「その他」に該当し、委員各自で適切に保管・管理・廃棄をお願いいたします。資料 2～5 については「取扱注意」とさせていただいております。厳重に保管いただいて、コピー等の複製、第三者への開示は御遠慮くださるようお願いいたします。それでは佐久間部会長、以後の議事進行についてよろしくお願いいたします。

＜報告書の論点検討＞

○佐久間部会長　御説明いただきありがとうございます。では、早速議題に入りたいと思います。本年 1 月 16 日に開催されました WG の参加者と概要を資料 1 にまとめております。それに関して第 3 回の議論を踏まえて、中田先生と田中先生には、市販後の性能変化のシナリオに関して追加資料の御提出を依頼させていただきました。

それから、12 月 5 日の第 3 回専門部会で委員の先生方から頂きました報告書の素案を取りまとめた上で、1 月 23 日に岡崎室長、副部会長の伊藤先生、中岡先生とともに報告書に更に追記していただくべき事項について整理、検討して御担当の各先生方にお願いをさせていただきました。その中でも書きましたが、データのいわゆる意図せぬ使用目的に対して、そのデータが適正なものかどうかといったような観点、それから独立性

の観点、そういったことが原則であるといったところを踏まえて考えていくということを議論させていただいてお願いした次第です。

そういう観点から、データベースのモダリティによってはそうはいってもなかなかできないという、現実問題として難しいという点もあると思います。一方、現行の捉え方の中でこういう課題があるといったところについて解決していきたいということで、お願いをさせていただきました。

市販後の性能変化シナリオに関する追加資料については、資料2ということで、報告書に更に追記いただいていますし、報告書に追記いただくべき依頼事項を含めた議論のための論点メモについては資料 3、「データベースのモダリティ固有の状況について」は資料4をそれぞれ御参照いただければと思っております。

また、再掲ということですが、12月5日の第3回専門部会において共有した報告書項目・素案については、改めて資料5に再掲しております。

まず、報告書の論点検討(3)に関する議論を進めていきたいと思います。つきましては、資料2を御覧いただければと思います。それでは、田中先生と中田先生、順番に資料2について追記をすべき点と、田中先生におかれましては、今日議論すべき点を書いていただいていますので、その辺りで少し進めていただけませんでしょうか。よろしくお願いいたします。

○田中委員　以前の議論で、同一患者の情報は市販後に用いないという、同様の情報が入ることによって、独立性とか、そういうものが阻害されるという意見が出ていたかと思います。これは非常に真っ当な意見だとは思うのですが、ただ、大規模な手術、外科手術であったり治療手技はともかく侵襲の少ない検査、心電図とかも入ってくると思いますけれども、患者が年次で定期的に受けるような検査の場合は、同一患者であっても時系列の変化・変遷がある以上、同一とするという必要は余りないのではないかと思っています。これら全て、画像というものに対して、単純な画像を学習させるという考え方にするのか、それとも付帯情報として、例えばここに病気と書かせていただきましたが、抗がん剤で治療して小さくなっていく、そういった変遷、あるいは変化が伴うようなものであったりすると、一定の条件であれば、同一患者でも使うことは問題ないのではないかと思っています。

もう1つ、これは内視鏡だけなのかどうか分からないのですが、やはり正常病変ですね、病変とのコントラストを持つ中で、正常のものを学習させないと、いい学習がなかなかできないということがあるのも現実でして、そういうことを考えたときに、できる限り多くの情報を得ようと

思った場合に、同一患者というくくりだけで駄目というのは、なかなか難しいのではないかということを書かせていただきました。

後で、それぞれのモダリティの所で議論が出てくると思うのですが、要するに同じ画像であっても、タグの付け方、病名だけが付いている、あるいは日時情報が付いているというようなものがあると、付帯情報によって同一患者であっても違う画像であるという考えで利用できるようなものがあるのではないかと。この辺の書きぶりは難しいと思うのですが、注意点としてそういうものを挙げさせていただいたということです。私からは以上です。

○佐久間部会長　ありがとうございます。では、先生方これについて何か御意見があればお願いいたします。

○伊藤副部会長　今、田中先生が最後に同じ画像であっても付帯情報の付け方によって違う情報として取り扱うという話をされたのですが、分かる部分もあるのですけれども、何かピンとこないところがあります。そこをもう少し具体的に教えていただけると有り難いのですが、よろしくお願いいたします。

○田中委員　ありがとうございます。1 つは、例えば潰瘍性大腸炎のような、治療によって画像が推移していく、そういうようなことがありますよね。重症化してから治療で良くなっていく過程というのがあった場合に、単に潰瘍性大腸炎となりますと、同じ患者だと、同じような写真しか集まってこない可能性があります。しかし、そこに病期、治癒期とか寛解期とか劇症期とかという、そういうのが付帯情報として付いていれば、同一患者であってもその画像というのは全く同じものではないということ、少し説明の仕方が悪かったのかもしれないのですが、同一患者だからといって同じ画像ではない場合もありますよねという、そういうことを触れさせていただいたということです。

外科手術とか、あるいは内視鏡でも大きな治療ということになりますと、同一患者で一生の間に 2 回も 3 回もなることはほとんどないと思うのですが、そういう侵襲の少ないもので経時的な変化を見るようなものというのが、やはり後段の議論になるかと思うのですけれども、モダリティによって違うものがあるのではないかという、そういうことで意見を述べさせていただきました。

○伊藤副部会長　ありがとうございました。同一患者であっても、何らかの治療介入や何らかの介入によって起こり得る変化を、変化として経時的に見る場合は、有用な情報として得る可能性があると、そういう理解でいいですか。

○田中委員　　はい、ここの議論ではそうだと思います。後ほどモダリティによっての違いという点では、特に工学系の先生方に教えていただきたいのですが、正常画像を学習するという扱いをどうするかによって少し違うかなと思いますけれども、今の議論の中では、伊藤先生がおまとめいただいた形で私の意見はほぼ網羅されていると思います。

○伊藤副部会長　ありがとうございます。

○佐久間部会長　今の議論なのですが、恐らく臨床的に見ると、同じ患者さんがどう変わっているか、病期とおっしゃいましたけれども、そういうことに関連した変化を見るということになると思いますが、それを少し工学的サイドから見ると、データの独立性というか、確かに同じ患者さんで、同じ場所なら、相関はあるのだけれども、病期が変わると相関性が落ちていきます。だから、そこの多様性というか、相関性が低いようなデータをちゃんと入れるということはあり得るのだという、そういう見方になるのかなと思います。その辺り、特に工学系の先生、どういう見方をして捉えていったらいいかというのをコメントいただければと思います。清水先生、どうぞ。

○清水委員　　田中先生、どうもありがとうございました。少し工学寄りの言葉を使って説明させていただきますと、恐らく先生がおっしゃっているのは、1 人の被検者の方の縦断的時系列データという言葉を使ったりするのですが、同じ人がどのように変化していくかというのをトレースしていって、その方の未来の状態を予測していくとか、あるいは現在の状態を診断するというような診断支援システムを考えた場合には、当然その方の過去のデータを使ってシステムを設計するというのはあり得ると思います。

一方で、横断的時系列データという言葉があるのですが、一人一人についてはある特定の時刻のデータでしかないのですけれども、それをたくさん集めて、集団として時間的に変化していくものを追っていって、時間的な変化をモデル化して、それからの逸脱を評価することで異常を検出するような、そういう横断的時系列データをベースにした場合には、ひょっとすると学習データにテストデータと全く同じ人がたまたま入っている場合と入っていない場合で、性能評価の結果にバイアスが掛かったり掛からなかったりする可能性がありますので、横断的時系列データをもとに設計する場合には、少し注意が要るかなとも感じました。工学的にはそのように問題を整理すればいいのではないかと思います。

○佐久間部会長　清水先生、貴重な御指摘を頂きましてありがとうございます。結果としては、意図する使用目的これがある人の将来を予測すると行った内容に関するものででであれば、そのような目的に対して見ると、今、田中先生

がおっしゃったように、同一の患者のデータを検討するということはあるだろうと。一方で、先ほどおっしゃっていたように、横断的時系列データは病気の進展の度合が違うから、そういうものによっていろいろ出てくるということがあった場合に、それを全体として見て捉えたときに、かなりのデータ量があってバラエティに富んでいれば良いと考えるのが妥当なのかという問いがあります。ひょっとすると、そこにある人特有の特徴が埋め込まれていて、それに対して応答してしまう可能性があるということかなと思うので、その辺りのリスクがあるかもしれないという、そういうことでしょうか。

○清水委員　補足をありがとうございます。おっしゃるとおりだと思います。科学的にはそういう心配もあるというような問題を指摘しておきたいと思います。ただ、それによって、使ってはいけないというようには禁止はしません。結局、これはｎの話も絡んでくると思うのです。テストデータが十分あれば、その辺りの問題も緩和されるでしょうから、複合的に問題が絡み合っているので、問題点の指摘というのが私もよろしいかと思います。

○佐久間部会長　多分その辺りを、書き方を注意して書いていくということだと思います。ただ、使ってはいけないということではなく、やはり使うシナリオとか使われる用途、そういうことを考えたときに、データの網羅性とか、それがパターンマッチングしようとしている事象だとか、そのようなことの特徴に従ってこれを使ってよいかどうかということ、あるいは使ったときの注意点のようなことを考える、そういうことを記述するということになるかと思います。他にいかがでしょうか。

○清水委員　これは後でもよかったのですが、せっかく今、データの独立性についての話が出ていますので、私が考えたことを少し口頭で述べさせていただきます。学習データとテストデータの独立性というのは非常に重要な概念だと思います。FDA もそれを大事にするというようなスタンスと理解しております。ただ、独立性の中にいろいろな種類の独立性があるということを、田中先生がおっしゃっているのだと思います。

例えば、時間に関する独立性です。同一患者であっても時間が十分たっていれば、起こるイベントというのは独立と考えることができる場合もあるとか、疾患の種類が違えば、例えば大腸がんとそれとは全く違う原発の別のがんは、それぞれ独立に起こる事象と考えてもいいでしょうし、疾患が発生する臓器とか部位、この話も確か田中先生の文書の中にあったかと思いますが、そういう臓器とか部位に関する独立性というのも考えて、そういう観点から許容できる場合もあるというふうに、独立性の

概念を時間、疾病、臓器、部位等に拡張して書くのがよいのではないかと思いました。

○佐久間部会長　ありがとうございます。その点は私もそうだろうと思います。結局、何をもって独立と考えるかというのは、多分シナリオを決定していかないといけないし、FDA の文書を見てもちゃんとそういうことが書いてありますが、どういう臨床条件で使うかということを考えてというのは常に出てくるので、その中で考えるということです。一方で、先ほど議論でいうとデータの独立性というようなことになるのだけれども、そこをつなぐ説明というか、そこをちゃんと書いておくのが、科学委員会のレポートとしては十分だと思います。今ですと、例えば臓器とか、時間、発生部位とか、いろいろな独立の観点というのか、そういうことがあるのだということを書いておくといったことかと思います。

○清水委員　どうもありがとうございました。

○佐久間部会長　続いて、中田先生、お願いできますか。FDA が出している市販後の性能変化に対する考え方のまとめだと思います。よろしくお願いいたします。

○中田(典)委員　前回の会で資料を提出するようにということで、比較的分かりやすいものはどういう資料があるかということを付け加えるためのものです。

資料 2 というのは、具体的に日米の規制のことについてです。FDA のほうは全般的な用語の説明と Pre-Cert の説明、これは FDA のほうでして、それと IDATEN のことが解説されているもの、それ以外のものは、実際の FDA のディスカッションペーパーの原本の文章とそれの説明を私のほうで少し加筆したものです。なので、補足的な説明だけなので、これについて特にこちらから加えることはありません。

資料 3 の最後のほうで、私のほうで 1 つだけコメントを書いておいたのが、4 番の画像診断装置側の問題、これも前回少し言ったのですが、ベンダーA 社の CT 画像で訓練したものをテストデータで審査された後に、今度はベンダーB 社、違う会社の CT 画像でテストした場合、AI 性能に変化が現れるということは当然あり得るということです。

それから、5 番目は対象症例の問題これはよく言われていますが、A、B、C、D の病院で収集された症例についてトレーニングしてテストしたのですが、承認された後に別の病院 E で症例を選定した CT 画像でテストした場合に、AI 性能に変化が現れてしまうとこういう市販後の性能評価にはいろいろな種類があるということについて補足させていただきました。

性能評価というのは、もう 1 つの宿題のほうでも確か書いたのですが、バイアスというのは科学的に存在するものなのですけれども、医学的にどうしても避けて通れないようなものです。それから、AI の研究開発に

は、根本的に薬剤の治験などとは異なった問題があって、これは症例数が違う、前向き研究ができない、そういう問題に起因するものだと思われるのですが、そういう分析手法の問題があるのです。これは、やむを得ないものがどうしてもあるのだという認識をしていただいて、承認を絶対的なものとして承認してしまって、それで終わりというのは無理だろうというのが、これらの資料分析等で分かってきたことなのです。

したがって、最初からそういうファジーな部分についてはやむを得ないものとして、一旦何らかの形で承認はしたけれども、その後、監視下に置くしか方法はないのではないかという流れというか、AI の研究を進めて市販するものも増えてきた中で、そのようになっていってしまっているという現状があるということを、委員の皆様方に分かっていただきたいなと思います。

○佐久間部会長　委員の先生方、いかがでしょうか。先生に付けていただいた FDA の原文と解説を合わせて見ていくと、面白いことに、変更がどう出るかということに対して、FDA の 3 ページの上の所に、「Categories of software modifications that may require a premarket submission include」と書いてあって、どのようなことで起きるかという話が書いてあります。多分、変更によって新しいリスクが出てくる、もともとあるリスクをもっと大きなリスクにしてしまうとか、そういう可能性とか、それから多分、リスクコントロールという区分になっていったのだと思います。

一方で、3 番目に「clinical functionality」とか「performance specification」ということが書いてあります。これと、IMDRF のキーワード、タームのディフィニションを書いたペーパーがあるのですけれども、その中のディスカッションという所に面白い議論が載っているのです。PMDA の岡崎さんから少し御指摘いただいたのですが、変化の原因というのをちゃんと書いています。変化を考えると、cause、effect、trigger、domain、effectuation ということは、5 つ考えなければいけないと書かれていて、コーズは多分、結局なぜ変更した理由になるのかということで、これは新たなデータに学習しましょうとか、ある機械モジュールを入れたから、そのような変更が生じる理由を明らかにするということです。

効果は何かというと、意図する使用の追加とか、適応の拡大とか、性能変化、入出力データの変化なども書かれています。それから、変更のきっかけというのは、新たなデータとか経験が出てきたり、定期的なメンテナンスというのが実はあるということです。それから、面白かったのがドメインの変化ということで、先生もおっしゃっていたここに書いて

ある均質なものと不均質なものに分けていて、普遍的に発生するものを均質に分類しています。全体に効いてくると思ったらそうではない場合もあることを指摘しています。例えば、同じあるクリニックに対応して変えていくという、ある病気に相当して変えていくといったような決定的なドメインの変化、あるいは病気が変わっていくといったような、そんなことが少し書いてあるので、多分そういう変化が起きる原因をちゃんと特定して、それによって出てくることを見ることが必要なのだろうなというのを意識しているのだろうと思います。これと、今のなぜこれを premarket review の段階で明らかにしなければならないかと書いてあることとが、多分リンクしているのだろうなと感じております。

それから、中田先生が書かれた資料 3 の絵で書いてある所がありますGood Machine Learning Practices、これは前々回ぐらいから出てきていることだと思います。そこにデータの質のことが書かれていて、それをちゃんとやってくださいと。その中で、AI に関して出てきたことは、多分モデルモニタリング、先ほど中田先生がおっしゃった実社会でのデータを見て、パフォーマンスを見て、それを戻すということをちゃんとやりなさいということがフィードバックされています。これは、実はトータルライフサイクルコントロールということです。ライフサイクルをコントロールするという考え方でやらなければ駄目だということを明確に言っているので、ここら辺が多分、中田先生がおっしゃっていた「やむを得ないだろう」というのは、そういうところかなと思っています。中田先生、恐らくそういうことですよね。

○中田(典)委員　そうなのですが、次の宿題の所でもう少し詳しく、なぜそれがやむを得ないのかというレポートを書きましたので、そのときに述べたいと思います。

○佐久間部会長　ここで、言葉として略語になってしまっているので、SPS、ACP というのは何のことかと申しますと、その下に書いてありますかね、SPS というのが SaMD Pre-Specifications、いわゆる医療機器ソフトウエアの何が今後変更計画されるのかということです。今の製品として、どれがどういう変更があり得るかということを、あらかじめ設定しておきなさいということを言っています。Specifications ですから、いわゆる初回の申請の段階で、どういう変更があるかということを明確にしておけということも書かれています。

Algorithm Change Protocol、ACP というのは、それをどうやってやるかということを明確にしてくださいというのを書いています。やはりこれも整理されていて、それを含めて Approved です。これは、実は IDATEN

と同じ考え方だと思いますが、それが出てきたものが入ってくると、Software modificationも入ってそれがうまくいっていれば、いろいろドキュメントだけでいけるようなケースがあるのです。一方で、初回に考えられたACPとSPSが少し違うといったときに、そこが軽微な場合は、Focused FDA review of SPS and ACPということをやるというふうに書いてあって、それから外れるもの、これで一番大きいのが意図する用途を変えるのが結構大きなところで、だんだんそれに近づくと、やはりpremarket reviewにいくということも書いてあります。

そんなロジックになっているので、多分そこに対して書かれているのだけれども、そこに対して、この中に今日これまで議論してきたようなデータの再利用の問題とか、それをどう入れ込むかということを今後議論していくことになるのかなと思っています。他の先生方、中田先生の御発表に対して何かございますか。それから、中田先生、私の解釈が間違っていたら指摘してください。

○中田(典)委員　解釈そのものは、この文書に書いてあるとおりのものを、どう解釈するかによって、人によって多少、微妙なところで差が出たりもするのでいいと思います。

○佐久間部会長　はい。多分そう明確なことは言っていないのです。だから、これに対してどうロジックを組み込むかというのは、結構重要なことなのかなという感じがしています。

結局、今まで議論していたAIというのは市販後は変わるのだけれども、野放図に変えるわけではないと。あるプランの中で変えていってくださいということを言っているのだけれども、そのプランも実はやっていくうちに変わってくる可能性があるという、そういうことも受け入れようねという形をしているのだと思うのです。それにしても、これを読んでいくと分かるのですが、一番簡単なのは、意図する用途も変わらなくて、ドメインも変わっていないので、性能が良くなるというのは、これはデータを集めてきて良くなったとか、そういうことで、あとは評価がうまくできていればよい。これは多分ドキュメントベースで承認されるであろうと考えられます。

ところが、意図する用途が変わるとか、そういうことだとすると、リスクが非常に大きくなってくるということです。

○中岡副部会長　この図というのは、基本的にはIDATENとほぼ同じだと理解しています。これまでにいろいろと話を皆さんから聞いたり、IDATENをどうやって使っていくかということを考えたときに、これはほぼ同じなのだろうなと。違うのは、日本の場合は510(k)がないだけで、そういう意味では、個人

的には日本でそういう 510(k)、DeNovo に相当するものを SaMD だけで設けるような余地があるかというところがポイントになるのだろうなと。今、見ていただいている資料のコメントを寄せられた後に、FDA が 2021 年に文書を出していますよね。FDA としても、例えばリアルワールドパフォーマンスをちゃんとフォローしないといけないとか、先ほどから議論が出ているバイアスをどうやって防ぐかということを、レギュラトリーサイエンスベースでちゃんと研究していかないといけないという、恐らく寄せられたコメントに対して返答はされているので、アメリカと日本が向いている方向は比較的一緒なのかなと思うところです。

もう一点は、Pre-Cert に関してですが、Pre-Cert の報告が去年出たと思うのですが、結局 Pre-Cert も長いことやっていたのに、アメリカでも結論が出ていないですよね。企業に対してそういう仕組みを要求するというのは、アメリカですら、若しくは企業なのか FDA のリソースの問題なのかもしれませんが、そういう特定の企業がいくら変えても OK だというような制度を設けるのはかなり難しいのかなと考えています。その辺りも、アメリカの最新状況と合わせてというか、向こうに合わせる必要はないのですが、日本は日本で、そういう事実も踏まえた上で、どういうところを要件として出していくのかということは考えていかないといけないのかなと思っています。

○佐久間部会長　ありがとうございます。その他コメントのある先生方はいらっしゃいますか。

○清水委員　聞き逃がしたのかもしれませんが、質問です。Software modifications というのは、再学習等による modification のことだけなのか、それともネットワークを大きく変える、例えば ResNet18 から ResNet101 に大きく総数を増やすとか、あるいは ResNet とは全く違うモデルを使うとか、そういうことも含んでいるかどうかという情報は何かお持ちですか。もしあれば教えていただければと思います。

○中田(典)委員　資料 3 にその訳を書いておいたのですが、ユースケースの具体的シナリオが書いてありますので、そこを見ていただければ分かるのですが、余りネットワーク自体を変えるのではなさそうですね。

○清水委員　その部分は、後で御説明いただけるのでしたか。

○中田(典)委員　いえ、後で説明するのはもう少し根本的な問題についてなので、ネットワークを変えるとか、そういう問題ではありません。ネットワークを変えてしまったら全く違うアルゴリズムになるので、多分また元に戻って、一番最初からやり直しになってしまうと思うのです。

○清水委員　そこはそうなのですね。分かりました。

○中田(典)委員　それはそうですよね。通常、常識的に考えればアルゴリズムが違う、ただしアルゴリズムを変化させたというのは、シナリオ 1B に書いてあるのですが、「データを追加し」と書いてあります。要するに、データを追加するから結果としてアルゴリズムが変わるということが書いてあるだけなので、アルゴリズムのプログラム自体を変えているわけではないような気がするのです。

○清水委員　　分かりました。

○佐久間部会長　少し読んでみると、そこは少し考えられているような感じがしていて、清水先生、例えば大規模なネットの構造が変わるのではなくて、ネットの構造は維持され、骨格は残っているのだけれども、細部が変わるということがありますよね。

○清水委員　ええ、あります。

○佐久間部会長　そういうことに関しては、何となく入っているのではないかなという印象は持ちました。

○清水委員　　そうなのですね。

○佐久間部会長　それをあらかじめ想定されているのであればということです。ただ、それを実際に、初回申請のときに認めるかどうかというのはまた別の問題であるかなと思うので、その可能性は否定はしていないのではないのかなという気がいたします。ただ、中田先生がおっしゃるとおり、全く違う構造のものが出てきたら、それは全体が違いますよね。だから骨格はある程度似ているとしても、もしかすると Premarket review がいるのかなという印象を読んでいて感じました。

○清水委員　分かりました。ありがとうございました。

○中岡副部会長　　もう一点補足でよろしいですか。アルゴリズムチェンジというのは、確かこの文書の中では具体的に定義されていないと思うのです。

○佐久間部会長　多分されていないですね。

○中岡副部会長　私の印象ですと、FDA のある意味さじ加減でどうにでもなるぐらいの感じです。ですから、佐久間先生が言われたように、これぐらいだったらいいのではないのというような総数が、清水先生は覚えておられると思いますが、前も AI の研究班の中でも議論になりましたけれども、1 層増える、2 層増えるというのが、大幅にアルゴリズムが変わることになるのかということが議論になったかと思います。その辺りのさじ加減が何となく FDA の中の裁量に入っているというような気がいたします。

○清水委員　　分かりました。

○中田(典)委員　　最近のはやりで、例えば AutoML のように、多数のアルゴリズムをパラメータによって最適化するというようなものがありますね。そのパラ

メータだけを変えるとテスト結果が変わってきてしまうのですが、そういうパラメータだったら、複数のアルゴリズムを使ってそれに重み付けをしたりして変えるので、重み付けのパラメータを変えるだけで結果が変わってくると。この程度の変化だったら、FDA は多分許すのではないかなと。そういうことだと思うのですが。

○清水委員　　どうもありがとうございました。

○佐久間部会長　それでは、次に行きたいと思います。資料 3 を御覧いただければと思います。資料 3 は議論したものの論点メモということで作りました。これをざっと見ていただいて、一部書き過ぎではないかという御指摘もあったので、そこについては修正いたしました。

それで、バイアスに関しては、利用手法のバイアスとか、人によるバイアスというようなところ、この辺りが実際少し特殊なように見えるので、その辺りをうまく表現していただけるといいかなと思って書きました。あと、多分バイアスそのものがあるということを常に頭に入れて、バイアスがないから大丈夫という言い方をしてしまうと、悪魔の証明になってしまうので、バイアスはあるのだけれどもこうだというような形で話ができるようなことが、多分現状の最適レベルではやむを得ないと思うので、そういうことを記述することになると考えられます。

それから、清水先生の御担当の部分ですけれども、これは先生が出された数学的な話があるのですけれども、極論すると、いわゆる訓練データと評価データこれが独立してあったとしても、それを良くするためにそれに近い訓練データを使うということは、その相関性が入った訓練データで訓練していることになってしまう。それがある意味で過学習を生む可能性があるのだということを言っているのだと思うので、その辺りの数学的なことプラス、もしかするとそれは書きすぎになってしまうかもしれないけれども、それらの懸念は多くの場合は持たなくとも良い、というような御説明をしていただけるといいのかなと思います。余りにもその危険性を強調するがゆえに、このような技術の研究開発ができなくなってしまうのは余りよろしくないと思うので、その辺りのバランスをどう書くかことがポイントになることを書きました。

それから、holdout とか、いろいろな言葉があったので、この辺りのところを少しまとめていただければいいかなと思います。

それから、鎮西先生のシミュレーションのほうで書きましたが、これは実は使ってはならないということを言っているわけではなくて、恐らくどう使ったらいいかという現状のレベルというのがあると思うので、その辺りの観点で書いていただくのがいいかなと思います。先生、数週間

前ですか、数値解析の海外での学会の情報とかを見てこられたかと思うので、その辺りも含めて、一方で医療データ、特に医療データは先ほどの議論を聞いていて思ったのですけれども、データにバラエティを持たせるといったときに、評価でやってはまずいのですけれども、学習の段階では人工的なものを使うというのは実は十分ありうる話かなと思っていて、そのときに、根拠のない人工的なものではなくて、根拠のある *in silico*-learning ということとの関連だと思うのですけれども、そういうことも出てくるかと思うのでその辺りかなと。これは後で議論したいと思います。それでは、鎮西先生何かありますか。

○鎮西委員　今、出ている部分では大丈夫です。

○清水委員　私の担当部分に頂いた御意見について、どういう方針なのかとか、それから少しよく理解できていないところがあるので、質問しようと思っていたのですが、5分ほど時間を頂いてよろしいですか。

○佐久間部会長　はい。

○清水委員　最初に頂いた御意見は、再利用のシーンについて最初の開発時に評価データを再利用する場合と、市販後に新たに別の学習データを使って学習させて、再び同一の評価データを使うケースです。

この2つ、確かにこの状況はありうると思います。フレッシュなテストデータを用いた評価結果に対してバイアスが含まれるという観点からは、恐らくこれは1と2で本質的な違い、例えば1や2に特有なバイアスはないのではないかと考えているのですが、もしあればコメントをお願いできればと思います。

それから、少し先に説明させてください。次の部分、この意義の部分です。医療データでは、アノテーションデータ付きのデータを集めるのは大変というのは、これは対応予定です。それから、先生が先ほどおっしゃっていた用語の Holdout、これも対応予定です。それから次、この部分がよく分からなかったのですが、「訓練に追加するデータの特徴が、評価データの特徴と類似する場合など、意図せず評価データに過学習されてしまう」「性能の向上のために追加した訓練データの特徴と評価データの特徴の類似性をどのような指標で評価するケールがあるか」、これは下の文章と続いている、同じ内容と考えてよろしいですか。

○佐久間部会長　私のほうの質問としては、いわゆるあるアルゴリズムを出すときに、使ったデータによる学習結果が適正かどうかを見るというものを見たときに、少し違う指標みたいなものを出されているという、例えばAUCを見るとか、そういうことをされていたというのがあったように記憶しており、このような過学習のされているのかどうかを判定する例を、確か先生が

おっしゃっていたような気がするのです。その辺り、機械学習に基づくＡＩを使う側としては過学習の有無を推測できないため，なんらか指標が必要ではないか。例えばこういう論文では、過学習が起きているか起きていないかを見るための指標として、こんなことを使っていたといったようなことがあればということです。もしなければ、いいのですけれども。

○清水委員　　分かりました。では、過去の資料を確認します。

○佐久間部会長　私の記憶では、この Holdout をなくすようなことをやるというアルゴリズムを、これが完全ではないけれども、こういうトライアルがされているということを紹介されたと思うのです。アルゴリズムを考えて、同じデータセットを使っても過学習を回避できるような手法が研究されていると思います。

○清水委員　　Dwork の話ですか。

○佐久間部会長　そうです。そのときに、その効果が上がったかどうかということを見るために、確か AUC だったか、そういうことを見ているということをしたと。私も十分によく読んでいなかったのですが、その辺りのところが、こんな考え方もあるということを示しておくといいのかなということを少し思ったので、それを書けるかどうかということです。

○清水委員　　過学習の検出方法ですか。

○佐久間部会長　そうですね。

○清水委員　　分かりました。過学習の検出方法はよく使われている方法がありますので書いておこうと思います。それから、この評価用データと類似のもの、あるいは同一被検者のデータを入れて確認するという点、田中先生から今日御指摘を頂いた点を踏まえてこんなふうに書こうかなと考えています。

これも小さくて恐縮なのですけれども、経時変化検出の際に同一被検者を学習とテストに利用する場合も想定されます。同一被検者でも疾患が異なる、疾患が発生した臓器や部位が異なる場合も考えられます。これは田中先生の文章をベースに、そのケースを書き込んでいます。そのため、独立性としては問題によっては時間に関する独立性、これは用語が適切かどうかは分かりませんが、仮に時間的独立性と呼びます。それ以外に疾患の種類、疾患が発生した臓器や部位に関する独立性、これも仮に疾患独立性、疾患発生の臓器独立性や部位独立性で、これらが成立した場合には、条件付きで適宜緩和する必要があるという点も追記予定です。

先ほどの説明と重複しますが、例えば同一被検者の縦断的時系列データ、同じ人をずっとトレースしているような場合ですが、それを用いた経時変化検出の場合には、過去の同一患者のデータを利用することは可とする。ただし、複数の被検者の横断的時系列データ、それぞれの患者さんのデータとしてはある時刻のデータで、それをたくさんいろいろなステージで集めてデータベース化して、それを使ってシステムを設計した場合、この場合には特定の被検者がテストに含まれているか否かでバイアスが掛かる可能性があるので注意が必要となる。このバイアスの大きさは問題です。対象疾患やシステムの狙いなどの種類やテストデータの規模にもよるため、影響の見極めには複数の要因をにらんだ総合的な判断となる。こういうようなことを書こうかと思っています。

それから、あとは摂動の話を頂いていましたので、これもそういう方法があるということを追記しようと思います。最後の点も追記予定です。

○佐久間部会長　ありがとうございます。大分クリアになってこの部分も重要なポイントだと思いますので、なにとぞよろしくお願いいたします。

それから、森先生はいらっしゃいますか。

○事務局（渕岡先端技術評価業務調整役）　森先生はいらしてないです。

○佐久間部会長　そうすると、別途これについては言及せずに議論するとして、これは非常に難しい概念なのだとだんだん分かってきました。それをうまく、例えばＮＩＳＴの文章等ではさらっと書かれているのですけれども、これを解説するとなると、難しいこともあるかなと思って、少しその辺りをどう原稿に書き込めるかについて、森先生とお話しておければと思います。

そこで少しドメインのバイアスというのも、実は医療環境の違いのの話もあって、結局データの品質に関してでも議論になりましたけれども、あるクリニック、ある病院の中のデータに対して最適化するのかという話と、全体の病院、日本全体に対して最適化する話なのかということ、例えばアジア全体でやるのか、それによっても随分変わってくるというところがあって、その辺りがやはり医療用のものを考えるときに結構重要なポイントになってくるのかなという気がします。医療上の状況、例えば同じ疾患を検出するのでも、病気がいっぱいくるところでの検出の感度と、一般の健康診断でやってくるところの感度とはまた少し違うでしょうし、その辺りというのをちゃんと意識したことが必要なのだというところが、市場のバイアスという、そこの話は結構ポイントかなと。

それでは、次は鎮西先生ですね。鎮西先生、数値解析の件、機械学習用データの作成とかありますので、追加部分があれば聞いていきたいと思います。

○鎮西委員　頂いたコメントでは、現段階では学習用に数値シミュレーションの結果を使うのはともかく、評価用データ作成には認められないという議論になると思われると断定されていたので、少しそこまで断定できないのではというのが私の今の意見です。

ただ、今載っていた注意事項としていろいろあるというのは事実なので、その辺りをどういうふうに盛り込むかだと思います。

まず、最初の文章は前からあった文章なのですが、その後ろに「数値シミュレーションにより評価データ生成」ということで、付け足してみました。現段階では、まだ克服すべき課題が存在し解決の道筋が不明であるということで、まだそこは未解決の問題がありますよということで書いておきました。

次に、2 つポツがあって、1 つ目のこちらのポツは、まず数値シミュレーションの信頼性とか信憑性といった話を 2021 年に出したコンピュータシミュレーション部会の報告書から引いています。そこでは、モデルがどういう成り立ちであるか、あるいはどういうデータに基づいてやったシミュレーションなのかという点でいうと、実験式的なモデル、あるいは動物とかの計測値で間接的なデータでやる数値シミュレーションというのは信憑性が違うということです。なので、数値シミュレーションそのものが臨床で計測、観察されるデータに代わるものとはみなされていないという具合に書いておきました。まず、これは事実としてしょうがないというか、まだ現状ではそうですということです。ただ、機械学習の評価というのは、治験や性能評価試験のように、end-to-end でやるものばかりではないのではないでしょうかというのは、これは実際にやっておられる先生方にお伺いしたいのですが、例えば特定の性能項目の評価をやりたいという場合、その項目に着目して数値シミュレーションを設定、実施して、V&V をやっていくことで、そういう結果も使える可能性はないのだろうかという点で、そういうことも書いておきました。

ただ、その場合、当然なのですが ASME V&V40 が言っているようなコンテクストオブユースとかそういったところ、あるいはそういった話と今度は機械学習で何を評価したいのかという項目との間に、ちゃんとリンクがないといけませんよという締め方にしてあります。これだと、まだ書き過ぎですか。審査側として、ここまで書かれるとかえって面倒だと

いうのがあればしょうがないですので。私は今のところ、こういう書き足しを考えているということです。

○佐久間部会長　いかがでしょうか、審査側から何か。

○岡崎プログラム医療機器審査室長　鎮西先生、お考えいただきありがとうございます。現時点での懸念事項が示されていて、実際の開発のときにその手法で開発をしようとしてきた場合に、その点について妥当な説明がなされるのであれば、乗り切れる可能性があるというふうに、先生はおっしゃっていたと思いますので、その可能性を現時点で否定できないのであれば、そういう方向性もありうるのではないかと思いました。以上です。

○佐久間部会長　ありがとうございます。他にありますか。そうですね、シミュレーションというのは、かなりどんどん精緻になってきていることは確かで、だけど入ってないこともあるのです。それがまさに目的なのでしょうけれど、そこはどれぐらい信憑性があるかということについてのコンセンサスができてくると、やれるわけです。これは V&V の考え方ですけれど、原子力発電所で事故を起こして実験するわけにもいかない。だから、そこで想定したものでということで、安全評価をしているという、実は破壊試験ができないので、そういうことが出てくるというのはありうるシナリオではあるのです。それが突拍子もないものが出てきても仕方なくて、目的に対して適切にくればということなので、それはもともとのコンピュータシミュレーションというものの考え方だったと思うのです。それに基づいて、信憑性があるデータをどう使うかということに関しては、確かにあるかもしれないという、このぐらいの書き方というものも確かによろしいのかなと思いました。

　では、先生、この辺りはよくご存じですので、加減をどうするかというのが非常に重要なところになってきますけれども、よろしくお願いいたします。

○清水委員　　よく分かっていないところもあるのですが、先生の文章の中の下から 5 行目の「特定の性能項目の評価を行うこと」というのは、例えば具体的に問題を考えてみて、例えば COVID-19 の CT を使った診断支援システムとかを想定してもよろしいのですか。そのときに、COVID のデータを集めるのが大変なので、人体の肺の CT をシミュレーション、あるいは病変部分だけをシミュレーションして、正常例に付け加えてみましたと。そして、それを使って評価するというようなことを想定されているのだと思うのですが、その場合に、特定の性能項目の評価というのは、例えばどんなものになるのでしょうか。

○鎮西委員　　私が考えていたのは、今の例だと、なかなか数値シミュレーションでという話は難しくなってくると思うのです。その１つ前のパラグラフで、演繹的に導出されたモデルというか、要するに実験式でないモデルで計算をしている例と実験式とでは、やはり扱いが違うだろうと。実験式がそもそもそれで本当にいいのかということに関して、誰かがデータを取って調べないといけないし、そのデータを取って調べていない範囲については、実験式というのは本当に合っているのかという話を、前の数値計算の部会でディスカッションしたところなのです。なので、今の話だとなかなか難しいかなと。

逆に可能であるとしたら、COVID のケースでうまくいくかどうか分からないですけれども、例えば CT の計測値に影響を与える何らかの物理的なファクターが加わっていて、そのファクターによってデータが歪められていると、そういった状況をあらかじめシミュレートしてあげて、データを加工するであるとか、そういったことは十分可能だろうと思うのです。

だから、ゼロからデータを作るというのはなかなか考えにくいわけですが、今あるデータを加工して何らかのそういう制約がかかった状況でも、きちんと機械学習はちゃんと見分けてくれますということを示すぐらいだったら、数値計算とかでも十分にいけるのかなと、そういう意味で書きました。

条件としては、この最後の 2 行に書いたコンテクストオブユース、数値計算をどういうコンテクストで使うのかということと、あとは、機械学習で評価したい項目というのがちゃんとつながっているということと、数値計算そのもののバリデーションがどういうふうに行われていて、機械学習で評価したい項目とどうつながっているのかという話を、きちんと説明してくれれば、数値計算で加工したデータ、その加工が 0.1 加えるだけなのか、0.9 を加えてほぼ 1 にするという話なのかというので、大分違うかもしれないのですが、シミュレーションだから駄目という話ではないのだろうなということで、こういう具合に書きました。

○清水委員　　分かりました。実は最近、アンダースペシフィケーションが問題だという方々が出てきて、その方々が言うのはストレステストをやりなさいと。データにストレスをかけて、機械学習のアルゴリズムの振る舞いをチェックしなさいと。そのストレスのかけ方が、例えば点広がり関数を変えてみたりとか、モーションブラーを加えてみたりとか、あと濃度値を変更するとか、そういうようなストレスが考えられるのですが、今おっし

ゃった数値シミュレーションというのは、正にその点広がり関数を少し変えてみたとか、モーションブラーとか。

ストレステスト、これはどのぐらいまで必要とされるのか、今議論がされていて、よく分からないところもあるのですが、もしそれが必要となると、おっしゃったように、それ全部、手持ちのリアルのデータの中から探してというのは、効率も悪いですし、現実的ではないので、数値シミュレーションでデータを模擬して評価するというのが、私もよいと感じました。

○鎮西委員　むしろこれは、例えば今のモーションブラーのような話になると、実際にそれを模したデータを実臨床で集めるというのは、相当効率が悪くなると思うのです。それよりはむしろ、これは純粋に物理学的な話になりますから、あるいは数値アルゴリズムで、CT アルゴリズムの話になってきますから、それをシミュレーションでやったほうがはるかに中立なデータが、合目的的なデータが取れるということになると思います。

○佐久間部会長　今の議論は前回の数値シミュレーションのところで、やはり物理学的原理とか、かなり演繹的にしっかり組み立てられていて、経験的なものが入っていないという、そういうところに関してはあるかもしれないです。少しその辺り、適切な例をまた先生方によろしくお願いいたします。

データベースのモダリティ固有の状況について、資料 4 の議論に入りたいと思います。

前回ですが、資料 3 にも書きましたけれども、大原則としては学習データと評価データは独立であること、独立の意味が先ほど少し広く取るべきだという議論になりました。それから、その製品が対象とする患者群の特徴についての統計的分布と評価データの統計的分布が等しいということが科学的に説明できるような形、そんな形で使えるようなデータを作るということが今後必要になるのでしょうけれども、その辺りについての現状というか、観点から少し見てみたいと思って、改訂をお願いしました。

それでは、順番にお願いいたします。まずは、伊藤先生、よろしくお願いいたします。

○伊藤副部会長　今、佐久間先生が言われた観点と違った観点で書いてしまったので、後でブラッシュアップが必要かなと反省しているのですが、まずデータベースの現状についてお話いたします。

手術動画のデータベースというのは、そもそも手術は同意を得ているのですけれども、プラスアルファとしてこういったものをデータベースに入れますかという同意がもう 1 つ余計に必要なわけです。そういった中で、

同意のためのサンプルバイアスが掛かる可能性があると言われていますけれども、データ活用に病院名などは削除されますよ、大丈夫ですよといって、なるべく収集を行ったという現状があります。

これは AMED 事業で 3 年間集めていましたが、今現状も動画のデータベースはオンゴーイングで集めています。今、4,000 を超えるぐらいで、やはり次の施設多様性というのもありますが、術式多様性とか、動画と一言で言っても、例えばここにある大腸・胃・肝胆膵・前立腺とこれだけのものがあって、疾患としての多様性とともに手術のやり方というのもそれぞれ違いますし、大腸の中でも術式が違うし胃の中でも違うという、そういった多様なものの中でこうやって集めているという現状はお伝えしなければいけないかなということです。

術式の多様性というところでお話しますと、もちろん御存じのとおり、こういった手術というのは、開腹の、おなかを大きく開ける手術というデータというのはほとんど取れなくて、最近では内視鏡手術、特に腹腔鏡手術ですが、最近日本では 14 領域に保険収載されたロボット手術もすごく増えていて、そういったものも術式の多様性として一方で集めなければいけないというのが現状だということです。

手術における 1 つの特徴というのは、やった手術の内容が実際の患者さんのアウトカム、アウトカムというのは例えばショートタームでいくと合併症だとか在院日数だとか、そういったものに関連しますし、長い期間だと要はがんが再発するだとか生きている生きていないだとか、そういう生存率ですが、そういったものに紐付いている中で合併症があるもの、大体皆さんこういうものというのは合併症のあるものは出したがらないのですけれども、リアルワールドとしては、ある程度こういう合併症の含まれたものも広く集めなければいけないということで、そういった提供をしていただいています。

その次、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と書いてあるのですけれども、これは医者側の多様な情報というものをかなり紐付けているというようなことを、我々の中で行っているということです。

次の収集時期については、我々が集めたのが 2020 年からなので、以前のものというのが余りないのですが、後でも説明しますが、手術動画というのは、時間的な時代的な背景によって、やられている内容というのが刻々と変ってくる部分が結構多いので、こういった意味においては経時的な、時代的な集める時期においても、こういうバイアスというか、そういったものがあるのではないかということです。

下のほうに、手術動画についての固有の状況についてということで四角の表にまとめました。手術動画特異的なものとして、間違っていたら後で御指摘いただきたいのですが、主に治療のために行う手術を動画で保存しているという中においては、消化器内視鏡のESTだとか、そういった治療ももちろんありますが、治療に特化するという意味ではやはり違うという部分と、手術動画に特化する少し違う部分というのは、異なる時期のデータが必要であるという意味においては、デバイスや手技が結構変わるので、こういったものがあると。あるいは、人に依存する手技が、特に同じ手術といっても、やられる人によって、またその人の体調とかももちろんありますし、剥離、展開などの手術手技、手技の判断などが、かなり多様に富んでいるデータだということがあります。

治療や検査が定型化されているかというのは△印になっていますが、大分日本では定型化がなされてきています。撮影記録に関わるデバイスの多様性としては、他のモダリティと同じようにあるだろうということです。それから特に一番下のデバイスの多様性というのが、ロボット手術の鉗子もそうですし腹腔鏡手術でも病院によって多様な機械が使われているという部分が、非常にこの手術動画の特徴だということです。

以上、次の3点です。治療の工程を全て記録した動画であるという点、手術は術者によって異なる判断、手技で行われる非常に多様性が多いという点、新たな手技や手法が出現してくることによって、時代的な変遷もあるということで、一方で違うと言いながら共通して例えば疾患や術式が違ったとしても、やはり上手な手術というのは組織に対する適切な緊張がかかって切るスピードであるとか切る深さであるとか、そういうジェネラルなものもその中にあるので、そういったものを一つ一つ考えながらデータを収集して、アノテーションして評価をしていかなければいけないというところがあると思います。以上です。

○佐久間部会長　何か御質問はございますか。アノテーションというのは、こういう連続した画像に対して、どういうアノテートをしている形になるのですか。

○伊藤副部会長　ある程度間引きした形になりますよね。例えば血管だとか尿管だとかといっても、動画は1秒間に30フレームなので、その30個を連続で塗ったとしても、余り多様な情報にならないので、ある程度間引きした、具体的に何秒とか、そういうふうにやって間引きして、アノテーションはしているということが多いです。

○佐久間部会長　見えているものの解剖構造とか、そういうことで。

○伊藤副部会長　はい。

○佐久間部会長　今後出てくる話としては、今、何の処置をしているのかみたいな流れですよね。その話というのも結構研究的にはあるのですが、その辺りはどうなのでしょうか。

○伊藤副部会長　おっしゃるとおりなのですが、デバイスとして今、何が出ているというアノテーションもそうなのですが、一方で今、何のシーンをやっているという手術というのは映画のようにシーンごとにやっていることが結構明確に最近は分かれているので、今どこのシーンだということをアノテーションしてフィードバックされることによってシーンが自動的に絞られてくるという部分においては、通常のデバイスのアノテーションとは違うアノテーション方法ですね。

○佐久間部会長　その辺りのことというのは、ある種のリスクマネジメントだとか、そういうことには結構効いてくる可能性があるので。

　そういうことが出てくるときには、恐らく画像と、今後でしょうけれども、バイタルサインのデータの場合とか、いろいろあるのだと思うのですが、共通的な部分はこうだということで、その辺りのことも含めて書いていただけると有り難いのかなと。

○伊藤副部会長　はい。ピントがぼけたような感じもしますが、何をアノテーションして、何を評価するかによってかなり変わるのです。

○佐久間部会長　その辺りを書いておいていただくと、自分たちが作るデータベースをどう作ったらいいのかという議論につながります。ずっと実は外科手術のデータベース作成維持に関与している方との話の中で，手術のいろいろなデータベースを作られているが，手術ににおいてどのデバイスを使ったかという記録は残していないという話がありました。このような記録は機器側からすると結構大変だよねという話をしていたこともあるのですけれども、何か漏れなくするにはどうしたらいいか。一方で、何でもデータを集めれば良いということになると，アノテーションをする先生のデータベース作りにおける作業が重くなるだけであって、その辺りはどうバランスを取ったらいいかということが課題だと思うのですが。

○伊藤副部会長　ただ、リアルワールドの使用頻度というのもある程度分かっているのです。例えば、超音波凝固切開装置だと HARMONIC が 7 割で何とかが何割でと。だから、ある程度本当に薬事的な部分でいくとリアルワールドをある程度反映した比率でそういう動画を作らなくてはいけないなというのもあると思います。

○佐久間部会長　そうですね。考えていらっしゃるような課題も書いていただくという形で。ありがとうございました。笹野先生、お願いします。

○笹野委員　　私からはバイタルデータということで、特に心電図を中心に書かせていただきました。大きな問題としては2つで、データに関するものと、アノテーションに関するものということです。

データに関する問題点ということなのですが、本日の議論でも、同一患者の複数データをどうするかということはずっと出ているのですが心電図や脈波信号などというのは、比較的簡便に記録できるので何度でも取れると。では、それらは同一データの再現を見ているのかという問題があるのですけれども、特に心電図などでは、日内変動や日差変動が大きいということがありまして同一被検者の複数回記録のデータというのは同一情報ではなく変動そのものがむしろ解析対象となることが出てくるという点が心電図というのが1つの波形でもあり、それを繰り返すデータとしては時系列でもあるという点における問題かと思います。ですので、これは長期的な治療介入前後ではなく、もっと短期的な問題ということになってきます。

それを考えると、b のほうの心電図データというのは、1 拍の波形が繰り返されているということになりますので、それをどのように捉えるかと。b に書いてある文章自体は、データ収集の話ではなくて処理の話のように見えますけれども本質的には、a と b は同じことで、心電図というのは1拍分のデータだけがあればいいのかというと必ずしもそうではなく、むしろ1拍だけで学習させると通常は精度は悪くて多数を集めたほうがいいというようなこともあります。ですので、これを考えると心電図というのはより長時間の変動も含んだ指標、変動の情報も含んだものが完全なその人の心電図データであって、今、我々が手にしているものは、その一部を抜き出しているものだけだというように考えられましてそれを 1 人1枚だけで独立というようにして考えていいかどうかというのは、また検討が必要かなという意味で書かせていただきました。

データに関する問題点の c ですが、これは例えば技師さんの電極装着部位のずれだとかということを書いています。これは施設間バイアスという中に含まれる話かとは思うのですが、あえて書いたのは心電図などのバイタルデータは今は広くいろいろな所で取れるようになってきて、ヘルスケアの領域との境目が曖昧になってきているということから、いろいろなデータが混入してくるだろうということで、そのときに、どうやって取ったかというようなことが必ずしも測定条件としてきれいに記録されていないことがありますのでそれを考慮したデータ収集が必要になってくるというようなことを問題点として挙げられると考えました。

2 番、アノテーションのほうです。アノテーションに関する問題点はいつも出てくるもので、1 つは、心電図の自動診断、目に見えるものを我々が判定してそれを再現させるというものは特にアノテーションの問題は発生しないのですが、発作性の疾患をどう予測するかというところでは、必ず出てくるのがコントロールはどういう人がコントロールなのですかということで発作性不整脈で自覚症状がないようなものというのは、100%のコントロールというのはあり得なくてこれはコントロールであるということを許容する基準というか、それをある程度示してあげないと、いつまでも生まれてから今まで 100%何もないのですねという悪魔の証明のようなものを続けなければいけないということがあります。

また、疾患群というのは、発作が記録されれば発作群、疾患群というように分類できるのですが、それは一方で心電図の記録の時期からプラスマイナス何箇月までの間に発作が記録できればいいのかという、そういう基準もまた曖昧でそこの基準のセッティングも必要というように思っています。

例えば、突然死を起こすような疾患の場合には、死んでいなければイベントは起きていないのだろうということでコントロールは、比較的明確にできる部分もあるのですが、そうではないもの非致死性かつ無症状の場合はアノテーションのものでは常に出てくる問題で、これは得られている心電図そのものには情報としては出てこないものですのでこのようなものをデータ活用のときには考慮しながら使う必要があるというように考えました。簡単ですが以上です。

○佐久間部会長　ありがとうございました。波形のデータというのは、長時間の時間的変動を見るといったようなものについてはそこの特性を踏まえてデータを取らなければいけないということを御指摘いただいたと思うのですが、アノテーションに関していうと非常に診断が難しいというかそういうものが存在するのだけれども今後そういうことを考えているものが出てきたときに、そういう限界があるのだということを理解した上で使うということになるでしょうか。

○笹野委員　そのとおりです。

○佐久間部会長　ありがとうございました。それでは、中田先生、お願いします。

○中田(典)委員　医用画像、放射線科領域とか超音波のバイアスについての文書です。医用画像については、大きく 2 つ問題があると。バイアスとか diagnostic imaging、医用画像についてパブメドで検索すると査読付き論文で、非常にインパクトファクターが高い論文について調べてみると、絞り込んで 2 つの問題があるということが分かったのでそれについて述べます。

まず、最初が研究開発方法の問題です。これは 2020 年の BMJ の論文なのですが、医用画像診断のディープラーニングの登録されている臨床試験全ての論文について調べています。2010 年から 2019 年の 6 月までの全ての臨床試験について調べていて、何を調べたかというとどういう研究方法をしているかという、この筆者は要するに無作為化試験について本当に無作為化をしているかということに興味があったようで、それについて調べました。そうすると、ほとんどないということでした。通常の臨床試験では、無作為化臨床試験が当たり前でこれによって臨床的なバイアスを下げる工夫をするわけですけれども、深層学習の医用画像を使った臨床試験では無作為化臨床試験が 10 件しかなくて、ほとんどが無作為化試験ではなかったということです。10 件の無作為化試験も、ほとんどが前向きではないということで、結局これによって臨床試験の方法論からすると医用画像の人工知能の研究というのはバイアスがもともと高いのだということが明らかになっています。ところが、実際にこれを厳密に、通常の臨床試験と同じように無作為化試験をしてしまうと例えばそれしか認可しないというようにすると AI 医療機器認可の障害になることは目に見えていますので、そういうものだということを分かった上で審査する以外に方法がないという、大きな問題が 1 つあるということが分かりました。

2 番目が、対象とする医用画像そのものの問題です。ここで参考にした論文というのは 2022 年のもので米国科学アカデミー紀要オンライン版に掲載されている論文です。『Implicit data crimes：Machine learning bias arising from misuse of public data』、これはパブリックデータのオープンにされているデータを使った研究というのが、そもそも適応と少し違う使い方をされているのではないかということを調べた論文です。

これによって分かることというのは、具体的にいうと例えば MRI でいうと、MRI はもともと使う画像そのものは DICOM データなのである程度人間が見て分かるようなデータとして AI に読み込ませるわけですが、それがウインドウとかレベルとか画像のフィルターとか撮影条件が全くばらばらだったりするわけです。その時点で、本来 MRI が持っている raw data とは違って、それによってバイアスがかなり掛かるはずだと。CT などはもっと著明で、64 列と 8 列では全然画質が違いますし、各種のウインドウレベルの条件とスライス圧とかそういうものによっても全然変わってしまいます。そういうものを考慮すると、もともとかなりバイアスを掛けることが可能になってしまうと。超音波などもメーカー間の画質の変

化、要するに画像処理を行っていろいろなフィルターを使ったりしますのでその時点でもうバイアスが発生してしまうという問題があるわけです。

もともと人工知能というのは、ある一定の専門家が見た場合どこのメーカーでもどういうベンダーであっても同じ診断を得られるような画像を使うというのが、人間の場合の評価の仕方だと思うのです。要するに、A社だったら分かるけれどもB社だったら分からないような画像を使っていたらおかしいだろうという話なので、それを共通化してどのようなベンダーであっても人間が判断できるものは、AI でもみんなが分かるものを使ってそれによって画像のバイアスをなくして、その後にそれぞれのメーカーが独自の画像処理を行ったものに対して、AI の製品を作るのは勝手なのですが最初はそれではいけないのだということだと思うのです。

ということで、どうしても清水先生や森先生がやられたような、コンピュータサイエンス的なバイアスを除くというのもとても重要なのですが、そもそも入力する画像それから入力した画像の研究方法これにバイアスがどうしても掛かってしまうのだということがあるので、この辺を考慮して審査する必要があるということを明確にした査読論文を2つ挙げさせていただいて記載しました。

○佐久間部会長　重要なことを指摘されているのかなと思います。確かにいろいろなアルゴリズムというのが入ってきて、論文を書くときはそれなりに論文の目的に合わせてデータを作成する可能性は十分にあるわけで、そのデータを用いた結果というのは、どこまで適正かどうかということについては少し批判的な部分も記載されているという、御指摘かなと思います。こういう論文が出ているということはちゃんとこういう反省が少し出てきているということですよね。

○伊藤副部会長　1点だけ教えていただきたいのですが、前半部分のランダム化の話なのですがランダム化する対象というのがどこをランダム化すればいいのかというところを教えていただきたかったのです。つまり、データをランダム化するのか、あるいは出てきたものに対して臨床的有用性を確認するために患者をランダム化することをしていないのかというのは、どちらなのですか。

○中田(典)委員　これは臨床試験のときと全く同じです。だから、通常は臨床試験の例えば第Ⅱ相だったら、第Ⅱ相を選ぶときに患者をランダマイズして選びますよね。それを AI でも同じようにテストするときにやれと。理想的には、それが臨床試験の在り方だと理想はそのように University College London の人は言っていますが実際には無理なのです。

○伊藤副部会長　ラーニングしたデバイスに対して、オンオフのランダム化をしてアウトカムを比較してくださいということですね。

○中田(典)委員　そうです。純粋な研究手法です。

○佐久間部会長　他にございませんでしょうか。田中先生は先ほどお話を頂いたということでよろしいでしょうか。

○田中委員　少し付け加えさせていただいてもよろしいでしょうか。

先ほどの伊藤先生がおまとめいただいたマトリックスを見せていただきたいのですが、非常によくおまとめいただいたと思います。何をここで言いたかったかというと「消化管内視鏡の動画」と書いていますが、先ほど触れていただいた内視鏡治療に関しては、手術動画と同じような問題点と扱いになるというように整理していただけると有り難いです。これからお話をするのは全て検査つまり診断のためのものの特殊性ということで、そういう条件で聞いていただけると有り難いです。

ここに「動画保存」と書いていますが、現実、動画保存で AI 開発をしている所というのは今のところは少ないのです。これは撮り方にもよりますが、1 秒 35 フレームの中で切り出した画像をやっていくときに非常に不正確な画像が出てくるということがあって、静止画でやっていることが多いということを 1 点注意させていただきたいというところがございます。

私の文書の中でアノテーションのことを余り書かなかったのですが、このアノテーションというのを、今、レギュレーションで決めるのは難しいと思っていて、最初、胃がんのディテクションのほうの AI をやるときに、矩形で囲むツールしかなかったのです。矩形で囲むツールでと言いますが、人体において矩形で囲めるような病変はないので多角形で囲むというようなツールをこちらで作ったわけですがそれが AI 開発において、正確なアノテーションがど動画像とともにの程度寄与するのかというところが、少し私には食い切れていないところがあります。病理などは結構高度な多角形のアノテーションをされていてその辺りがアノテーションに関して難しさを触れるのはいいのですが、規定してしまうとここも問題が大きいのかなというところで、あえて文章にはアノテーションを記載しませんでした。

3 点目です。私が書いた文章で、私が一番言いたかったのはタグ情報をどういうものを付けるかということによって、使い方が違うのではないかということであって、それが間違いだということでは全くありません。先ほどバイアスの話がありました。確かに機器や機種によって違う画像処理をするようなものがあるのは事実です。だからこそパラメータを付

けておけば、タグ情報でこれはNBIだこれはこういう処理をしたというのがタグ情報で付いていればそれだけでグルーピングできますからそれによってバイアスを除くことができると思うのです。つまり、画像にどのようなタグを付けるかというところの議論が必要なところで、付帯情報があればあるほど困難ではあるのですが、正当性というか、そういうものが得られると思いますので、そこを強調しておきたいということです。

ただ、問題なのは、先ほど伊藤先生が触れられましたが手術を目的とするような治療などでは個別同意が取れることがあって、そのときに「AIに使います」という同意というのは取りやすいのですが、検査の場合は難しいところがあります。パラメータが数多くなると臨床情報が画像に付くことになるので、これが機微な個人情報であるというようにいう方々がおられて、こうなると同意がなかなか難しいということもあるので、そういったところを、今、伊藤先生におまとめいただいたマトリックスの中にうまく落とし込めればいいのかなと思いました。以上です。

○伊藤副部会長　様々なモダリティで、これが一番いいとは私は全然思っていないのですが、幾つかの項目の中で、今のデータベースで共通するものと違いというのを、1枚で明確にできるようなものを作ると分かりやすいかなと思いました。

○田中委員　伊藤先生、これはすごくいいと思います。分かりやすくなると思うのです。いつまでもばらばらでそれぞれの分野で話していると議論が散逸してしまうので、こういうマトリックスをお作りいただけると非常に分かりやすくなると思いましたので、御礼を申し上げたくてお話をしました。

○伊藤副部会長　ブラッシュアップは必要かなと思います。ありがとうございます。

○佐久間部会長　今、御指摘いただいたところで、消化管の治療の部分は動画像とともに、それから付帯情報、デバイス情報、そういう何をしているかという点ですが、それを付けるのは難しいところもあるのですが、付帯情報があるほうがより価値の高いものになる点等を御指摘いただきました。また、付帯情報があると使い方も様々に変わってくるということです。ありがとうございました。佐々木先生、お願いします。

○佐々木委員　伊藤先生におまとめいただいた表に尽くされていると思うのですが、病理の場合には標本作成もマニュアルで行われていましてHE染色の染色方法、表に書かれているように染色液のメーカーも違えば自動染色装置も出ていますけれども染色する時間のプログラムも違うということで、肉眼で見ても明らかに違う。（またプログラム開発に使う）画像がその上に書いているように、通常はバーチャルスライドスキャナというデジタル画像装置を使うのですが、そもそもこの装置が日本で販売されている

のが9社ありまして、全てDICOM規格になっていなくてフォーマットがばらばらなのです。結局、データがそもそもかなりデバイスが多様になっていて、病理学会がAMEDから研究費をもらってプログラムの開発をやったのですがなかなかうまくいかなかったと。諦めたわけではなくていまだにバリデーションをして一生懸命やっているのですが、5つぐらいの施設で非常にいい結果が出たというものを、6つ目に持っていくとなかなかいいデータが出ないという、感度、特異度ともに惨憺たる結果になるというようなものが出ていて、そのデータのいわゆるデバイスの多様性をどのように克服していくかというのが、集めたデータに対してどのようにやっていくのかというのが課題かなと思っております。簡単ですが、このようなことを中心に書いています。

○佐久間部会長　ありがとうございます。今の点、先ほど中田先生が出されたFDAのフローの中でも、対象機械の入力という入力データというところの変更というものに入ってくるということで、ある機械に入ってきた入力データというものを限定してある程度示すということはできるのですけれども、それが入ったときに今言ったようなことが出るということだとすると、その辺りも考えておかなければいけないということが現実の世界の難しさである。また、先生はすごく苦労されたことがあるので、ある意味ではどうするべきかということを書いていただくことがいいのかなと思います。

これで大体全体の議論は終わりですけれども、全体を通して、先生方、御発言いただいていない先生方も含めて、御意見があれば頂きたいと思います。

○伊藤副部会長　佐々木先生に一言だけなのですが、私も本当に思うのですが、病理ほどばらばらで、画一的なプレパラートまでの作成方法がないと結局無理なのではないかと思うようなところもあるのですが、その辺についてはいかがですか。余りにもばらばらすぎて先生の言われることはそのとおりだなと思ったのですけれども。

○佐々木委員　これは質問も兼ねてなのですが、例えば集められたデジタルデータをデジタル加工したもので、いわゆるAIの医療機器プログラムなどを開発するというのは、例えば薬事承認のときに許されたりするものなのでしょうか。

○佐久間部会長　今のはデジタル画像処理で、例えばこういうことでしょうか。いろいろな機器の差だとか、そういうことはあるけれどもそれを見たときに、ある一定の標準的なものに画像処理をすることによって標準化する正規化

するようなプロセスをして、それでもってやるということは許されるかということでしょうか。

○佐々木委員　はい。いわゆるレアデータではなくて、全て加工データを使ってその加工データを使って出てきた結果を、もとのデータの診断と合わせるような操作をしながら医療機器プログラムの開発を行うということは許されるのかどうかという。

○佐久間部会長　それも処理の１つだという形で体系に入れていくという考え方はあると思います。それが余りにも多様になってきたときに、どのようなことが起きるかという。そこは、いわゆる画像変換みたいなことがあって、滅茶苦茶な変換ではないのだけれども、施設間の差だとか機器の差だとかそういうことを吸収するような処理があり，そのアルゴリズムでロジックで正規化が行われることも変わるかもしれないし、それこそ機械学習を使っている所がやってくるということもあるかもしれないけれども、その正規化するような形、データにして、開発した技術の評価を行うことがある。正規化されたデータに対してやったことというのは許されるのかという質問ですね。

○佐々木委員　はい。なおかつ、その出てきたものの判定をもとの診断と同じように合わせなければいけないのでそこでまたアノテーションを付けるというか。

○佐久間部会長　これは、恐らく MRI はあるかなとは思っていて、MRI はまだ機械の差がすごく大きいのです。定量 MRI を研究している先生から聞いたのですが、なかなかそこがうまくいっていないという。CT は比較的しっかりしている標準化されたデータとなっているのですけどその辺りを画像に対する AI 医療機器に関する議論に入れようというのはあり得るストーリーです。そのようなことが出てきたときにどうするかということに少し似ていますので、コントラストが違っているとかコントラストを正規化してより見やすくするとかいろいろあると思うので、そのようなデータの前処理による影響をどう考えるかというところかなと思います。他の先生方、この辺りはどうでしょうか。鎮西先生、どうぞ。

○鎮西委員　恐らく同じような話がいろいろな所にあり、あとは、少し分野は違うのですが、例えばウェアラブルデバイスで活動量計があって、あれをいろいろな会社のものを付けてみるとみんなばらばらの結果が出るというのは、みんなが知っている話なのですが逆に、それをどうすれば統一的に扱えるかという話をやっている人たちも結構いるのでこれからそういう意味でのデータ標準化といいますか、データの加工の仕方というのは１つのトピックになってくると思います。

なので、今の話も正にその一部になり得るのでこの種の話は「バイアス」という言い方は適切ではないと思うので、別の呼び方をするようにしたほうがいいのかなと思いました。以上です。

○清水委員　1 点コメントです。前処理をして画像を標準化するというのは、ほとんど全ての診断支援システムが内部で行っている処理となりますので、この議論は機械のみが使う場合と、医師がその画像を診断根拠として例えばカルテに付けるとか、そういう話は分けて議論していただいたほうがいいかと思います。

○佐久間部会長　以上で大体よろしいでしょうか。資料 5 に今まであったものがまとまっていますけれども、またこれは本日頂いた議論を載せていただいて最終の報告案にまとめていただくように、そろそろ収束するという形にしていきたいと思いますのでよろしくお願いいたします。本日の議論で議論はまとまってきたかなと思います。ただ、全体のところで、先ほど中田先生が書かれていたスキームの中に、この議論がどう取り込まれるのか、本日の議論がどのように関連するのかという点とそこはまとめ始めていますので、その辺りは中岡先生、伊藤先生と議論させていただいて、また先生方に早めに出して、プレビューをしたいと思います。事務局から他に何かございますか。

＜その他＞

○事務局（渕岡先端技術評価業務調整役）　次回の専門部会ですが、4 月 4 日(火)の 14 時から 16 時の開催を予定しております。詳細などについては、追って御連絡いたします。

また、3 月 2 日(木)の 10 時から 12 時には、第 4 回専門部会 WG を予定しておりまして、御参加いただく委員の先生方を御案内いたします。佐久間部会長、中岡副部会長、伊藤副部会長、佐々木委員、笹野委員、清水委員、田中委員、鎮西委員、中田典生委員、森委員、横井委員でございます。事務局からは以上です。

＜閉会＞

○佐久間部会長　本日の専門部会はここまでとさせていただきたいと思います。長時間にわたり、ありがとうございました。また、WG も本日の議論を受けて報告書をまとめたいと思います。よろしくお願いいたします。