機密性２

第２回ＡＩ専門部会

日時　平成２９年３月２３日(木)
１７：００～１９：００
場所　ＰＭＤＡ会議室２１～２５

＜開会＞

○光石部会長　それでは定刻も過ぎておりますので、第２回 AI 専門部会を開催させていただきます。本日はお忙しい中お集まりいただきまして、ありがとうございます。まずは、事務局から委員の出席状況の報告と資料の確認をお願いいたします。

＜委員出席状況確認及び資料確認＞

○事務局(江原)　まず、委員の出席状況について申し上げます。14 名の委員のうち、12 名の先生に御出席いただいていることを御報告いたします。

次に、配布資料の確認をお願いできればと思います。お手元の資料ですが、議事次第の下に資料一覧がありますので、そちらをお手元に御確認をお願いできればと思います。資料としては、資料 1～5 まであります。それから、参考資料となっております。あと、机上配布としまして「委員名簿」を配布させていただいております。

資料ですが、資料１は報告書骨子(案)、資料２は宮野先生の講演資料、資料３は清水先生の講演資料、資料４といたしまして、PMDA の医療機器審査部の講演資料、資料５といたしまして、今までの論点整理となっております。お手元に無いものなどがございましたら、事務局までお願いできればと思います。

最後に、資料の取扱区分につきまして御報告申し上げます。資料 1、資料５につきましては、「その他」といたしまして、各自で保存、適切に管理をお願いいたします。それから、資料４は「取扱注意」とさせていただいておりますが、こちらにつきましても「その他」の取扱いとさせていただければと存じます。それ以外の資料の２と３につきましては「取扱注意」、すなわち、厳重に保管して、コピーなどの複製、第三者への開示は御遠慮いただければと思います。以上です。

○光石部会長　ありがとうございます。過不足は大丈夫でしょうか。それでは、今日は森先生が初めて出席ということですので、簡単に自己紹介をお願いしてよろしいでしょうか。

○森委員　名古屋大学の情報基盤センター長を務めています森と申します。どうぞよろしくお願いします。私は清水委員と一緒に長い間、機械学習等を研究室でやっておりまして、それを使った医用画像処理システムの研究をずっとやってきています。よろしくお願いいたします。

＜第１回専門部会（平成 29 年１月 13 日）の振り返りなど＞

○光石部会長　森先生、どうもありがとうございます。

それでは、第１回の専門部会は親委員会との合同開催でして、松尾先生にAI全般について、それから、須田先生に自動運転の御講演をいただいたところです。親委員会も含めまして、これまでのAIの議論を資料５にまとめておりますので、資料５を見ていただければと思います。本専門部会にてこれから記載する報告書のまとめ方の案である資料１ですが、これは講演の終了後に討論をいたします。また、参考資料として、厚生労働省のAI活用推進懇談会（AI懇談会）の資料を配布しております。大江先生が構成員をされておりますが、何か留意点等がございましたら、お願いいたします。大江先生、説明をお願いします。

○大江副部会長　厚生労働省にAI活用推進懇談会が設置されまして、12月の下旬から議論をしています。前回の会議の資料の一部を今日、参考資料として出していただいております。これはAIによる診療支援の安全性、責任の考え方、医師とどう責任を分けるのかというようなこと、それから、有効性の確保などについてどういう考え方を持つ必要があるかというようなことで、十分な議論はこの懇談会でもなされていませんが、こういう形の資料で前回出されているものですので、参考までに御覧いただきたいと思います。

＜議題１：AI（人工知能）等に関するご講演と意見交換＞

＜①IBM Watson for Genomicsの活用経験から（東京大学医科学研究所　宮野悟教授）＞

○光石部会長　ありがとうございます。参考になるような内容も議論がされているということですので、こちらで既に議論されたものは、場合によっては本専門部会でも参考にさせていただきたいと考えています。

早速ですが、本日は3つの講演を用意しておりまして、その後に、報告書作成に向けた全体討論を行いたいと考えております。

最初の講演ですが、「IBM Watson for Genomicsの活用経験から」というタイトルで、東京大学医科学研究所のヒトゲノム解析センターの教授であられます宮野悟先生に御講演をお願いしております。宮野先生は大変著名な研究者ですが、元は数学者であり、ゲノムデータベース解析、システム生物学による遺伝子ネットワーク推定といったようなバイオインフォマティックスの世界をリードする先生です。宮野先生、どうぞよろしくお願いいたします。

○宮野教授　御紹介いただきましてありがとうございます。宮野でございます。ヒトゲノム解析センターと名前が長くて、皆、ゲノムセンターと呼んでいるのですが、もっと進化が進むと、ゲーセンになると思っております。ちょっと粗野な言葉も交えて話をさせていただく失礼をお許しください。

「Artificial Intelligence 人工知能」という言葉が使われて、略してAI と呼ばれていますが、私は、個人的にはあまり好きな言葉ではございません。この「IBM Watson for Genomics」は、ニューヨーク・シティにある New York Genome Center でベースの開発が行われたものです。それを使った活用経験では、「人智の増強」(Augmented Intelligence)というのが私どもの感想です。学習・推論する辞書のようなものです。人(専門医・研究者)を replace するものでは決してないものです。それと、データに依存しているので、データがなければ AI は完全無能であるということも強く思い知らされました。IBM は「Cognitive Computing」という分からない言葉を使っておりまして、これも、私は使わないようにしております。

なぜ Watson for Genomics を医科学研究所の臨床シークエンス研究の中で導入しなければならなかったかと申しますと、私どもは 2011 年から始めたのですが、人の目での変異の整理、解析、解釈(キュレーション)は極めて大変であるということを経験してまいりました。これは、エクソーム解析で全遺伝子の変異を調べて、これは血液ですが、その中から 1,477 の変異が見つかってきた。それから、キュレーションを始めた医科研の血液腫瘍内科の専門医による自験例でやりますと、1,477 から、最終的にドライバー遺伝子を 20 個ほど出して、薬剤標的変異を見いだすのに 2 週間以上かかっております。病院の業務も行いながらですので 24 時間というわけではありませんが、この 2 年間、大変自虐的なことを何度も何度もやっていただいております。

それと同時に、なぜこういうことをやってきたかといいますと、Watson for Genomics がどれくらい有効であるか、どんな改良点があるのかということを見いだしていくということも、私たちの研究の目的でした。これを使いますと、ほぼ同じ到着点に、大体 10 分以下でできます。これも後でお話します。

学習データは、これは 2015 年 7 月 1 日に導入されたのですが、その当時は PubMed のアブストラクト 2,000 万件以上、それと、1,500 万件以上の特許情報、それと、体細胞変異のデータベースである COSMIC、これは今、400 万以上の体細胞変異が 2 万 4,000 ほどの文献に紐付けされているものですが、それと NIH の Pathway のデータベース。それとアメリカでは統一されている治験情報、それと ClinVar、これはヘルスや病気に関するバリエーションのデータベース、NIH のものです、それとあと、Elsevier などの出版社との契約による Golden Standard のフル論文など、これらを使っております。技術としては、機械学習の技術です。コンピューターの発

達で、昔は「不可能」なことが簡単にできるようになったというのが実感です。IBM はディープラーニングのために NDIVIA 社と協業を始めたなどというニュースも出たかと思います。

一方、自然言語処理は人手による専門領域のコーパス作りが背景にあります。つまり、人手が要らないというわけではなくて、人手が大いに要り、専門領域の人が必要です。Watson のヘルスのほうの開発には、MD が 40 人、薬剤師が 80 人関わっていると聞いております。この場合はがんの Genomics ですが、応用領域に合わせた細やかな推論方式や使いやすいインターフェースの作り込みが必要です。それと、良質のデータの利用が肝だと感じております。この技術だけでできるものではないと思いました。今、Watson for Genomics には、The Cancer Genome Atlas というところのデータや、そのほか、様々なデータがバックグラウンドで使われております。

あと、その絞り込みです。今、ちょっと概略を申しましたが、最初、変異の集団があります。変異の文献情報、機能予測などを通し、変異の絞り込み、そして標的変異、Actionable な変異と呼んでいますが、それの抽出、そして、変異とひもづいた薬剤情報の読み込み、そういったことをやって絞り込んでいくというプロセスが中で行われております。

Watson for Genomics は、あくまで支援ツールと認識しております。できることは、がんの発症に関連した遺伝子変異の候補を指摘すること、変異に対応する分子標的薬の提案、これは FDA 承認薬、日本の治験情報とか承認のものではなく、アメリカのものしか使えない状況になっております。FDA 承認薬・アメリカでの臨床試験情報、それと、薬剤耐性情報などの提示がされます。変異に対応する情報として、関連疾患や遺伝子機能などを提示してくれるというものです。

病気を診断することはできません。臨床情報・変異情報をもとに患者さんの病態を解釈し、治療法を指南するなどということも、もちろん無理な話です。実際の解釈・判断は、Watson から提示された内容を見たキュレーター/医師が担っております。専門知識がなければ Watson が提示する結果は意味を成さないというものです。

東大医科研のがん臨床シークエンス研究の体制ですが、患者さんから同意を得てシークエンスをして、スパコンで解析し、そのときに、変異を見付け出すのに私どもは Genomon というのを開発して使ってきているのですが、それで出てきた変異、これは構造変異なども含みますが、それを検討するわけです。ここから患者さんにつなげるときには、ここに医師が入る。ここに大きなギャップがあったのですが、Watson でスムーズに

いくようになったというのが私どもの感想です。ここから直接、患者さんにいったり、この機械から患者さんに何かが行ったりするということはないというやり方で進めてまいりました。

これは、同じことを繰り返しますが、説明と同意取得、検体採取、ライブラリ調整・シークエンス、これはWETのチームと申しますが、それとインフォマティクス、DRYのチーム、そして支援ツールがあって、解釈する人がいます。これは人間、生身の専門医の方です。そして、この総合したものをTumor Boardの中で、色々議論をいたします。そして、その結果を臨床カンファレンスのほうに戻し、臨床介入が行われるというのが、私どものところで行っているものです。絵に描きますと、Pre-WET、WET、DRY、そして解釈者がいて、担当医に回して、それを支援ツールのエンジンで回していくということで、全エクソーム、全遺伝子解析ですが、現在、5日でこのプロセスが走るようになりました。最初に始めたときは20日間ぐらいかかっていましたが、今、全遺伝子解析において5日間で走るようになりました。医師の人たちは、本業と兼業しながらやっているわけです。

Before WATSON、これは果てしない人海戦術でした。After WATSON、キャッチアップできたかなという感じがいたします。そのお話をしようかと思います。

ビフォアワトソン、全ゲノムシークエンス解析です。これは古川洋一先生がされたのですが、大腸ポリポーシスです。[redacted]この患者さんでは全然原因が分からなかったため、全ゲノムシークエンス解析をやって、これは大変な作業ですが、目で色々調べていって、その中で古川先生とうちのテクニシャンの者がブラウザで見ていっているのですが、「先生、ここのところがちょっと薄くなっているのですけど」と言うので調べますと、「これ、欠損しているのじゃないか」と。そこでロングレンジができるシークエンスでやりますと、10kbのdeletionがあって、ここに実はAPC遺伝子の制御を担っているプロモーターの1つが入っていたというわけです。これはその後、また文献を手作業で調べるのですが、それを調べると、個々の異常が、プロモーターの欠損が、このポリポーシスに関係しているという文献が見つかってきたというものでした。これには半年以上の時間がかかりました。実際に全ゲノムシークエンスのannotationをやって返している、例えばハドソン・アルファ研究所というところがありますが、これはgerm lineベースですが、約90日で返すというプロセ

スで走っているそうです。

アフターワトソン、絶望感からの再起というものです。

[redacted]

原因不明の患者さんの全ゲノムシークエンスをやりました。そうすると、250 万の SNV が上がってきました。古川先生が Watson に、これをそのまま、変異のファイルをアップロードすると 30 分ほどで 100 ほどの遺伝子が絞られてきて、この色の濃さはエビデンスの強さです。それで着色されているものです。これから、どんな薬が有効かというので、これは PDGF receptor αというものですが、それをターゲットにしたレゴラフェニブと、これは大腸がんなどにも使うものですが、それが上がってきている。あと、ペリフォシンという、これはフェーズ 3 の治験が行われているものです。それと、オフラベルの抗がん剤がごそっと上がってきています。こういったものを見て、医師は患者さんの病態を捉えようとするものです。クリックするだけでいろいろなことが分かるわけです。

同時に、NCI の Pathway のデータベースがありますので、それにマップされています。これは、クリックすると、どういう遺伝子で、どんな変異がこの患者さんには入っていてというのが、全部サーベイできるようになっているものです。例えば、先ほどのレゴラフェニブがターゲットにしていた PDGF receptor αというものは multi Kinase Inhibitor ですが、こういった所に機序している。それとか、ペリフォシンは、一応、AKT1 の阻害剤ですが、この患者さんは AKT1 には変異が入っていないのですが、上流にあるボスですね、FLT3 とか、JAK2 とか、PI3K とか、PDGFRA もそうですが、そういったものに変異が入っていて、ボスが、働け、働け、働け、働けと暴言を吐くので AKT1 も働かざるを得なくなっているので、この方に上司の暴言が伝わらないようにする阻害剤が有効であろうというので提言されてきているものです。

こういったものが、薬についてはこうやってクリックするだけ、フェーズ 3 の治験の結果はどうだったかとか、そういったものが、こんな感じでクリックすることで出てくるようになっています。ペリフォシンについても同じです。

血液の場合は、コントロールのためには口腔粘膜を germ line として使います。腫瘍のほうは骨髄液などを使っていっているわけです。スピードが問題です。診断が困難な血液疾患は全遺伝子解析をやったものですが、AML-NOS、not otherwise specified、よく分からないというものですが、これの全エクソーム解析をやり、144 のものが最初にスクリーニングで上がってきたのです。これは構造異常も含まれています。それを Watson に放り込みますと、こんな感じで、少ないものですが上がってきます。並行して、先ほど自験例がありましたが、専門医による人海戦術的解釈プロセスを実施して、同じようなところに行くかということを見ていきました。これは Watson が提示した Actionable なもの、米国で Actionable なものですが、承認されているものはないということでした。ただ、フェーズ 3 のもの、フェーズ 1 のもの、そういったものが上がってきているということです。同じ FLT3 をターゲットにしたものやキザルチニブなど、こういった薬剤がクリックするレベルで上がってきて、医師及び研究者はそれを見て次の自分の判断に使うというものです。同時に、このキュレーションのプロセスで人海戦術的なことも同時にやっていまして、それで最終的なレポートを臨床カンファレンスのほうに持っていくというものです。

時間の関係で早飛ばしにさせていただきます。ここに今、Watson とキュレーションの 2 つだけしか書いておりませんが、ほかのデータベースやエクソマイザーなどを使って検討して、複数のものを走らせて見ていっているという状況です。

これが血液腫瘍内科のものです。全患者数は、これは 2015 年 7 月から去年の 11 月までですが、113 人。そして、シークエンスの種別は色々あるのですが、その中から、最終的には、Informative なものは 73、Actionable29、Action を取ったものが 8 症例ありました。承認薬が 4 症例、臨床試験中のものが 2 症例、その他のもの、オフラベルのものが 2 症例という格好で医科研で実施されております。マニュアルで人が人海戦術でやったものと、Watson がやったものは、解釈は一致しません。オーバーラップはあります。というものです。

ただ、その中で感じたことは、例えば急性骨髄性白血病再発というように病理で診断されたもの、ゲノムのシークエンスをやってキュレーション、Watson をやって出てきたものは、Action の候補は、こっちは急性なのですが、慢性骨髄性白血病承認薬剤が上がってきました。結果としては化学療法選択や、これは Philadelphia 染色体陰性のもので、提示されてきたものは、最終的にたどり着いたのは Philadelphia 染色体陽性のも

のというような形で、病理の診断とは違ってきているものを経験してきております。Precision Medicine を見据えた遺伝子/ゲノム別の薬剤適応が必要だということを改めて感じさせられました。

それと、ELSI に関してですが、AI にインプットするデータに関して、どのようにデータの品質を規定・評価するか、こういったところが問題です。最初、Watson が導入されたときに血液腫瘍内科のものを試してみますと、妙な答えしか返ってこなかったのです。どうしてかということを調べていきますと、血液のほうでトレーニングがあまりされていなかったと。IBM は、北米を除いて初めて日本に入ったのですが、血液でやるところというので医科研を選んだのだと思います。それで、血液で学習をどんどんやってくれということでやりますと、変なところもありますが、3 か月ほどで普通に使えるような段階に成長していきました。それと同時に、膨大なデータをどうやって確保するかという問題があります。先ほど申しましたように、データがなければ、AI は完全無能です。

それと、AI の技術を用いたシステム使用者の範囲に関してです。これは医療に限らないと思いますが、ゲノムの変異の情報はもう民間を通して自分で得られるわけです。こういう AI のサービスが普通に提供されれば、患者さんは自由に使えるような環境になる。そういった社会的イシューが起こってくると懸念されています。「消費者直販型」の DTC 検査などがありましたが、それと同じようなことが、もっと深刻な状況で出てくる可能性があるのではないかということも感じました。

それと、AI を介して得られる結果に関連してですが、国内未承認薬・治療への患者のアクセスの在り方として、こういったことが分かってきました。ここの ELSI のところのスライドは、医科学研究所の武藤香織先生と一緒に作らせていただきました。それと、実際にありましたが、予期せぬ結果(incidental findings)は普通に出てまいります。

あと、AI を用いた薬事規制上の位置付けです。「AI=プログラム+データ」なのですが、これが薬機法の対象になるかというのが、Watson for Genomics を導入するにあたって大きな課題になっておりました。これは IBM の本社のコンプライアンスのトップのほうから止められました。医科研でやることはこういうことだということを丁寧に説明し、臨床の先生方にも同席していただいて、インタビューをする中で了解が得られて、要は、臨床研究支援の研究であるという形で承認が得られて、使えるようになったわけです。例えば同じ変異情報ファイルを入力しても、1 か月後には提示される結果に変化があります。それは、データが増え、学習しているためです。こういった現実があります。

あと、これがほぼ最後ですが、医療ソフトウェア規制に関する米国の動向です。PMDA の方々も詳細に御検討されていると聞いておりますが、21st Century Cures Act というのが、昨年の 12 月 13 日にオバマ大統領が最終的に署名をして動き出しました。この中で、医療関係の機関の経営支援ソフトや電子カルテなど、これまでも非医療機器とされていたものに加え、以下のものは医療機器ではないことを明確化するということが、この Cures Act に書き込まれております。私どもが Watson でやっていたことが、ちょうどこの①から③に書かれているようなもので、患者個人の医療情報やその他の医療情報(論文など)を表示・分析し、医療関係者に診断、治療等の支援又は推奨(recommendation)の提示を行い、且つ、医療上の判断を下す際に当該推奨を最初から当てにすることのないように、医療関係者がそれらの推奨根拠を独自にレビューすることができるようにしているもの。ただし、画像診断情報や診断機器からの信号を分析するものを除く。すなわち、画像の解析ですが、そこのあたり、清水先生がされると思いますが、画像の解析などはこの範囲ではないと思います。

日本では、御存じのように、患者の体重等のデータから麻酔薬の投与量を容易な検証ができない方法により算出し、投与を支援するプログラム。これは医療機器という範疇です。一方で、添付文書に書かれてあるものをそのまま提示するようなものは非該当というような格好になっております。こういった 21st Century Cures Act に沿ったような考え方が日本でできると、こういう医療支援ができるかなと考えました。

この 18 か月で学んだことですが、最後、AI は専門医の判断を支援する検索ツールとして有用であると、AI の解釈と専門医の解釈は必ずしも一致しないことを認識しております。同時に、全遺伝子を調べないパネル解析の限界を思い知らされました。

それと、最後の青字部分ですが、ビッグデータといいながら、患者さんの病態を理解しようとするには、1 人の患者さんを理解するには、あまりにも基礎となるデータが、今、私が提示しましたデータが、データと知識が不足しているという現実があるということを認識しました。少し長くなりましたが以上です。

○光石部会長　宮野先生、どうもありがとうございました。10 分弱ほど、ただいまの講演に関する質疑応答をいたします。委員の先生方、いかがでしょうか。

○森委員　名古屋大学の森です。データを入れるときには、先生方のほうでデータを入力しているのですか。それとも IBM に依頼して入力してもらっているのですか。

○宮野教授　私どものほうで入力いたします。ウインドウにアップロードのところがあって、そこにファイルを読み込みます。

○許委員　許です、ありがとうございました。最後のデータと知識があまりにも不足しているというのは、先生が扱われた疾患が希少疾患だからなのでしょうか。

○宮野教授　そういうことではありません。がんの多様性と言いますか、ヘテロ性が非常に大きく、パネルで調べても候補となる薬はない。ドライバーとなるもので根拠の強いものはないということがありました。希少疾患であるからというわけではなく、例えばこのプロセスの中で、世界で4例目、日本では初の血液のディスオーダーが見つかったりしています。そういう意味で、非常にまれなものであっても、普通はつかまえてくるのが難しいのですけれども、パッとつかまえてくることができます。こういうのが XIAP という、IAP という遺伝子なのですが、それにコネ変異があったというのが見つかったりしております。

○大江副部会長　ありがとうございました。実際にやってみたときに、治療方針の決定に変更なり影響を明らかに与え得るような情報提示が Watson からなされる割合というのは、投入した症例のうちの何割ぐらいだと現時点では考えられるのでしょうか。

○宮野教授　ここに出しているインフォーマティブの意味ですが、病理診断をやったものが、なるほど自分たちの病理診断もちゃんと正しかったというものと、それがひっくり返るものがあります。きちっと数字は出しておりませんが、時々レポートが出てきます。感じとしては3割ぐらいがひっくり返っています。例えば、造血幹細胞移植を本来はすべきだった、というように最終的に診断が付いたものが、実はケモセラピーで治療が始められていた。それでシーケンスをやって出てくると、これは違うよねというので、造血幹細胞移植に切り換えたものがあります。また、まだ急性骨髄性白血病は発症していないのだけれども、骨髄形成症候群に関係する遺伝子の変異がボソボソ入っている、ただ、病理ではただ貧血という状態です。そのフォローアップで、普通だったら1年後に来てくださいというのが、1か月後にもう一回来てくださいという形でやるような、治療方針が変わったというのが普通にあります。

○光石部会長　似たような質問になるかもしれないのですが、次のスライドで、「Watson とマニュアルの解釈は一致しない」と書いてあるのですが、これはうまくいっていると理解していいのか、それともそうではないと理解していいのか、これはどういう理解をすればよろしいのでしょうか。

○宮野教授　難しいです。私のメッセージは、AI に頼りきるなということです。そ

れと21st Century Curesにも書いていますが、出てきた結果をそのまま信じてアクションを取るなということの教訓になるかなという形で用意いたしました。答えにはなっていませんが。

○鎮西委員　2つお伺いします。1つは、東大医科研を受診される患者さんが、患者群としてはどれぐらい一般的なのかという話はいかがでしょうか。

○宮野教授　どちらかというと、他の病院で分からなかったという患者さんが紹介されて来るのが多いと聞いております。

○鎮西委員　2つ目は、今のこういう状況が現在あるとして、データをもっと集積していけば何かしらコンバージするものなのか、はたまた人間のほうが追いつけない状況になってくるのか。例えば30年後になると、そもそもお医者さんが地道にそういうものを調査するという訓練をどうかすると忘れてしまう可能性もある。これから後どうなっていくのでしょうか。

○宮野教授　例えで説明して恐縮ですが、Google検索をしない医師はもういないと思います。それと同じで、こういう支援ツールのようなもの、知識にアクセスして、しかも効率良くできるだけ確度の高いもの、且つ、重要なことは、それを医療従事者が自分でレビューして確認できるプロセスを提供しているということ。そういうものが当たり前になってくるのではないかと思います。食べログなどのグルメサイトだと、本当においしいのかどうかというのは最終的に行ってみないと分からないというのはありますが、写真や色々なコメントを見ます。そういうところが、より科学的な根拠が強いものを辿って、辿り着いていくようなことになる。それが効率的にされるようになると思います。

○鎮西委員　今度は、そのエビデンスになるペーパーのクオリティが問題になってきます。例えば20年後に、お医者さんはどんな感じでペーパーを書くことになるのでしょうか。結局ベースになっているのがAIだとすると、鶏・卵的になってくる。

○宮野教授　私の今までの経験だと、医師は絶対に存在すべきで、診断は医師がやり、診断の責任は医師にあるので、機械側のこういうプログラムにあるとは思いません。

○許委員　今、我々の領域の論文では、一例報告などというのは単なる経験というので、むしろ前向き、無作為、割付試験というのが非常に重宝されています。それで、僅かの差を統計的に処理しています。一例一例のデータベースを蓄積していくほうが、エビデンスとしてはそういう前向きのものよりも強くなる可能性はないでしょうか。

○宮野教授　それは、あるかと思います。今は数が少なくテスト的にやっている段階です。これが日本人のベースでたくさんのものが出てくると、精度、確

度は上がってくると思います。これは科学的な根拠があるわけではありません。

○許委員　今は多くのジャーナル、日本のジャーナルもケースレポートは非常に評価が低くなって、採用率がうんと落ちてきています。これに対して警鐘を鳴らしてみたほうが良いのではないでしょうか。

○宮野教授　そうなのです。ですからこのケースレポートを探すというのは、現場では非常に大変なのです。それが、これでは全部一応、引っ掛けてきてくれていて、Watson からのリコメンデーションの中には基本的に含まれると考えていいかと思います。ケースレポートがあるというのは非常に。実際に最後の詰めはケースレポートがないかと探すわけです。

○光石部会長　ありがとうございます。まだ質問はあるかもしれないのですが、時間の制約もありますので次に移らせていただきます。宮野先生、どうもありがとうございました。

＜議題１：②医用画像処理における人工知能の応用の現状と課題（清水委員）＞

○光石部会長　次は、東京農工大学の清水先生に、医用画像処理における人工知能の応用の現状と課題について御講演をいただきます。AI の中でも、医用画像は実用化が最も近い将来に実現されるのではないかと言われていますが、その課題についても共有していきたいと思います。清水先生、よろしくお願いいたします。

○清水委員　御紹介ありがとうございます。東京農工大学の清水です。いただいたタイトルで発表いたします。本日の発表の背景なのですが、宮野先生からありました Watson のようなものは、自然言語が入力の主な対象となっています。前回 1 月 13 日に東京大学の松尾豊先生からお話がありました深層学習、これは画像だけではなくて音声、それから化合物反応予測も対象になっています。本日の話は、医用画像にフォーカスを当てます。人工知能の中でも、前半は特に深層学習、松尾先生のお話のおさらいになるかもしれませんが、医用にフォーカスを当てた話だと考えてください。後半は計算機支援診断(CAD)、この研究・課題についてお話させていただきます。

発表の目次なのですが、まず前回の粗筋をざっと話をした後で、トップカンファレンスにおける深層学習関連発表の状況について説明をします。医用の分野の深層学習について再考します。その後に重要なこと、私の私見を述べさせていただいた後で話を変えて、画像に基づく CAD のインパクトと想定されるリスク、それから現状と課題の話をさせていただきます。

まずこの2つについて話をします。おさらいになりますが、松尾先生が第三次 AI ブームの話をされました。そこでビッグデータと計算機の進歩、更に深層学習におけるいくつかの工夫が組み合わさってブームが来たという話でした。そこでは革新的な方法論として、画像から自動的に有用な特徴を抽出するという仕組み、これは Google の猫が有名ですが、これも例に挙げて話をされていました。それから、コンテストでの突出した実績も見逃せない重要な動向です。近いうちに「眼を持った機械」が誕生して、機械・ロボットの世界でのカンブリア爆発が起こるのではないかという話でした。

医用も含む画像解析のトップカンファレンスでは、具体的にこれは読みませんが、一言で言うと演題数はうなぎ上りに増えている状況です。「踊る」というような表現もされていますが、こういう形容がぴったりくるような状況でもあります。私が考える深層学習についていくつかコメントを述べさせていただきます。

Google の猫というのは、おばあさん細胞仮説を計算機の中で実証したという意味で、すごいインパクトがある話でした。おそらく同じ分野にいる人は、かなりインパクトを持って受け入れたというか、発表論文を読んだ記憶があると思います。それから、コンテストの成績もすごいものがあります。でも、それだけで本当に眼ができるのでしょうか、という疑問も一方で存在することも確かです。

もう一度、松尾先生の資料を精査させていただくと、非常にバランスの取れた資料になっています。例えば、体や病気の仕組みについての「総合的な理解」がないと診断できない問題では、医師を超えるのは無理ですという話も書いてあります。

ちょっと考えてみましょう。例えば医学的知識を持たない、つまり、解剖・病理をまったく知らない人に、全身の CT とラベル、このラベルというのは正常とか異常、それだけを大量に見せて、果たして画像を理解できるようになるかと考えると、それはできないと考えるのが当然だと思います。

したがって、今の技術では同じ問題を機械で解かせようとしても解けないということになります。しかも、計算機の中のネットワークというのは、脳に比べてはるかに貧弱です。今はもう少しこのコネクション数は増えているのかもしれませんが、大体 10 万倍、コネクション数、変数の数が違いますので、まだまだ人間の脳には及ばないレベルです。

そうは言っても、Neural Network というのは、昔から非常に研究者を引き付ける力があって、それはヒトの脳という実例があるという意味で、

筋が良いアプローチではないかと考えるという考え方は確かに存在します。いずれ人と肩を並べる、あるいはそれ以上の性能の人工知能の主役の1つとなるようなことは、ほぼ間違いないでしょう。

ただ、これが今かどうかというのが問題になっているのだと思います。今回のブームをもたらした技術をもう一度おさらいして、どのタイミングで本当にブレイクスルーが来るのかというのも考えてみます。

これは、今の深層学習で用いられているネットワークの例なのですが、こういうものがあります。頭文字には大体「Deep」が付いているのですが、あえて今は外してあります。実は、ほとんどのネットワークは過去に既に提案されているものです。今回の新しい点は多層化です。層を増やして深くした。それによって、従来はこういう問題がありました。一つ一つは挙げませんが、これらの問題に対して、例えば初期値は Pre-training をすれば良いよねとか、転移学習をすれば良いよねというような話があります。その他のものについてもいくつか工夫があります。ただ、どれも経験的に開発された技術であるという側面は否定できないものがあります。

従来との違い、これも松尾先生のところにまとめてあったのですけれども、ネットワークは本質的には従来と同じで、データ数と計算機パワーは桁違いに増えています。学習のための技術的工夫もあります。それらが組み合わさってデータ数、計算パワーが桁違いに増えて、いくつかの学習アルゴリズムの工夫によって性能が向上したというのが現状だと思います。楽観的に見る方は、これにプラスして、特徴が自動で抽出できるようになったことも踏まえて、これから大きく広がっていくと主張されます。逆に批判的に見る人は、違いはこれだけでしょう、というようなことを言う人がいるのも事実です。

この話を踏まえて、今の深層学習の限界を考えてみます。コンテストは非常にインパクトがあったのですが、良質なデータが十分与えられた同一条件下で、深層学習以外の技術で本当に同じ性能が達成できないのかというのは、実は誰もちゃんと検証していないのです。例えば、コンテストの成績の違いというのは、研究機関独自の学習データの違いにも依存しているという話もあります。

それから、これがちょっと大きい話と言えば大きい話なのですが、多くの処理はブラックボックスで解析困難です。第二次の AI ブームが去っていった重要な理由がここの点にあります。もちろんいくつかの優れた研究は出てきているのですが、まだ完全に解決されたとは理解されていません。また、三層のネットワークで、原理的には任意の識別境界が表現

可能という話というのは前世紀に証明された話です。多層化によって性能が向上する本当の理由というのは、実はいまだに不明です。証拠は少しずつ集まりつつあるのですが、理論的な解明はまだ不十分と考えています。

それから、ネットワークの構造やハイーパーパラメータの決定というのもいくつかの研究から出てきているのですが、依然として黒魔術、ノウハウの塊だということ。それから人の脳と同規模のネットワークにするためには、接続数を 5 桁(10 万倍)増やす必要があります。現代の方法論の延長にその超大規模ネットワークを学習できるアルゴリズムはあるのでしょうかという疑問もあります。

この話を踏まえて重要なことを考えてみました。今のはちょっとネガティブに聞こえたかもしれませんが、深層学習は間違いなく今の AI ブームのキープレーヤーというかスタープレーヤーであることには間違いはありません。そのキープレーヤーは万能ではないということも、今の話から見えてきます。そのことを踏まえて、得意なこと不得意なことを精査して使い分けることが重要ではないかと考えています。

例えば、得意なことはどういうことなのかというと、よく言われているのが大量の画像とラベル組が収集可能な問題です。それ以外に画像解析の研究を見ていると、こんな条件も見えてきます。対象の特徴が、画像中の位置に依存しないとか、対象は画像の中央や全体に広がっているという条件があります。

どんな例があるかというと、顕微鏡画像中の細胞検出とか、細胞の良悪性鑑別、あるいは CT 像や X 線像上のびまん性疾患や腫瘤の良悪性鑑別というようなものが得意そうな問題として見えてきます。

ここで少し話を変えます。鳩による病理画像診断と良悪性鑑別のスライドです。これは 2015 年にアメリカで行われた真面目な研究の話です。鳩に良性と悪性の画像を見せます。問題としては他に、乳房 X 線画像の良性腫瘍と悪性腫瘍も見せます。鳩は、これが何かはまったく理解をしていません。この黄色と青のボックスのどちらかを突っつくと、それが正解であれば（黄色いボックスが良性で、青いほうが悪性であるというように人のほうは紐付けていますが、鳩はまったくそれを理解しない上で、どっちかを突っついてその答えが正解になると）餌がもらえるシステムです。

それで 2 週間訓練をしました。

個々の鳩の性能は 1 が上限だと思ってください。ここに書いてあるような性能なのですが、4 羽の鳩を組み合わせると 99％の認識率が出ました。

この論文は多くのことを示唆していると思いますし、考えなければいけないこともたくさんあると思います。

例えばこの問題がどの程度の問題なのか。医学的に簡単な問題なのか、あるいは難しい問題なのかということも考えないといけません。深層学習に対してもこの実験は示唆に富んだことを教えてくれていると私は理解しています。今の深層学習で、まったくネットワークに病理の知識、それから解剖の知識を入れない状態で、ブラックボックスのままで学習させるということは、見方によってはこういう研究と似たような危うさを抱えていると考えます。

したがって、深層学習の研究というのは、そこから抜け出す必要があると思います。例えば、理論的な解析が進んで、ブラックボックスではない状態に早く持っていく必要があると思います。誰も、あの鳩あるいはそれに類するようなブラックボックスに診断されて喜ぶような人はいないと思いますので、そのような方向も大事だと思います。その一方で、楽観的に見ようと思えば、これは単に重要な画像特徴が何であるかを人間は気づいていないのだけの話で、実際に問題が解けてしまうのであれば、こういう問題はどんどん深層学習とかそういうもので解いてしまえば良いのではないかというような考え方もあると思います。いくつかの考察すべき点はあると思いますが、こういう研究もあります。

それから不得意そうな医用画像処理の問題について話をします。画像とラベル組が少数しかない希少な疾患の診断というのは、深層学習にはもちろん向いていません。解剖や病理に関する高度な知識、例えば臓器の形状や、その統計的変動に関する知識が必要なもの、病気や手術により輪郭が変形したり、コントラストが低く解剖学的知識なしには認識困難な例が不得意な問題です。最近は、治療法が非常に進んでいるので、手術を何度も経験して、臓器が変形したような方でも、ちゃんと社会復帰されていますから、実際の臨床ではこういう問題が解けないといけないのです。この種の問題というのは、高度な医学的知識を入れないといけない、プライヤーを入れないといけないので、今の深層学習ではなかなか難しいのではないかと思います。あくまでも「今の」という前提は付きます。これからどうなるかというのは、もちろん変わっていくところであります。

不得意なことも、既存の技術と融合させることで解決できることも多くあると思います。実際これをIBMは始めているのですが、従来は自然言語のみを対象とした人工知能で、これに画像も突っ込んでもっと精度を上げましょうとか、他の機械学習の技術を組み合わせましょうとか、これ

は橋爪先生がやられている多元計算解剖学などの数理のモデルを使うというような別のアプローチとの組合せも考えられる。これからは、未来の話も含めて考えたいと思います。

今までは、現在のところから見える話をしたのですが、これからは深層学習の研究だってどのように変わっていくか分かりません。最良のシナリオを考えてみました。深層学習の研究が、理論的側面からも進み、理論的裏付けを持ってアルゴリズム・ネットワークが登場して、内部の解析はブラックボックスではないという状況です。様々な問題に対して、どれだけのデータを要し、どのようにネットワークを設計して学習すれば良いのかということが明らかになる。そうすると、まさに「眼を持った機械」が登場して、あとは松尾先生の後半の資料のとおりになっていくのだと思います。

一方、最悪のシナリオも考えられます。再び冬が来るかもしれません。それでも、おそらくDeepCNNとかいくつかのネットワーク・アルゴリズムというのは、その嵐をくぐり抜けて生き残るのだとは考えています。大事なのは、次にまた春が来るということです。それがNeural Network単独か、あるいは他の技術との組合せかというのは分からないのですが、大事なのはそのX dayに備えることではないかと思います。

その備えなのですが、まず私が一番注目するのはデータです。医用画像でデータを気にするのはこの2つの問題です。正解ラベルの収集が難しい。正解ラベルは、急いで作ってもらっても、誤りが入ってしまっては、なかなか学習の効率が上がらないのです。こういうあたりも将来は変わるかもしれないのですけれども、良質であるということは重要なポイントです。あとは個人情報保護法との関係です。

ここから10分ぐらいを使って、CADのインパクトとリスクと、現状と課題の話をさせていただきます。まず、CADについて定義します。計算機が画像を解析し、例えば腫瘤の位置、良悪性などについて定量化した結果を医師に提示するシステムで、これが世界初のCADのシステムです。マンモグラムの中から腫瘤、あるいは石灰化を検出するシステムで、まず医師は普通にマンモグラムを読影します。その後、横に付いているボタンを押すとコンピューターが解析をした結果が、モニター上に重畳されて表示されます。その結果を見て、医師は必要であれば自分の解釈を変え、必要がなければそのままレポートを作るというシステムです。もちろんメリットは、見落とし削減によるこれらのメリット、それから拾いすぎ削減によるこれらのメリットがあります。使い方によっては読影時間や人件費削減などということもあるかもしれません。

使い方について紹介します。主に3つの使い方があります。1つはFirst readerと呼ばれている方法で、1次スクリーニングとしてCADを使う方法です。実際にはこういう使い方をしているものはありませんが、理論的には考えられますということで書いてあります。

主には2つ目のSecond readerの方法で、画像に対して、まず診断医が読影する。CADも同時に読影しているのですが、診断医がボタンを押したタイミングでその結果を表示して、場合によっては解釈を変えるというやり方です。先ほどの例だと、このreaderというのは箱が2つありますけれども、実態は1つになっているケースです。

3つ目は、Concurrent readerと呼ばれている方法で、読影と同時にCADの結果を表示するというものです。もう最初からCADの結果がモニター上に表示されているケースです。

ここからはリスクの話をします。CADシステムはソフトウェアと、それを搭載するハードウェアからなりますので、こういう電気的安全性うんぬんの話も出てくるのですが、本日は関係ないと思いますので、このスライドはスキップします。

まず、誤診による色々なリスクが考えられます。見落としのリスク、これは先ほどのインパクトの裏返しです。それから拾いすぎによるリスクも、先ほどのインパクトの裏返しになります。それから法的なリスクですが、CADがSecondかConcurrent reader、かつ最終責任は読影医が負うというような前提でしたら、法的リスクがCADの開発者や販売者などに発生するとは考えづらいと思います。しかし、将来は、ひょっとするとFirst readerというものもあるのかもしれません。そうすると、リスクも生じてくると思います。

最後にCADの現状と課題です。現状は、まず米国から説明します。米国では、高い乳がん検診受診率、それから医療報酬が認められたという話、検診の読影方式は単独方式であったということ、医療の訴訟が非常に多いということがあって、1998年に先ほどのマンモグラムCADがFDAで認可・発売されて、それ以来様々な種類のCADが認可されてきました。これは2007年までですが、2007年までででもこれだけあります。マンモグラム以外にも、胸部X線像、それから脳もありますし、大腸CT像等々のCADが認可され、発売されています。

ただ、一方で有効性に関する疑問も時々出てくるというのも事実です。これは2007年にFentonという人が出した有名な論文です。有効性はないということを大々的なスタディで書いたものです。ただ、この批判に対しても、たくさんの批判論文が出ていることも事実です。現在は、どち

らかというと有効であるという論文のほうが多いので、現在も米国ではCADが使い続けられています。

国内では、ここに書いてあるようなものが薬事で承認され、2014年からはソフトウェア単体の認証も可能になりましたし、色々なガイドラインも整備されてきていますので、今後国内では拡大予想と考えています。また、米国を中心に、先ほども説明したとおり相当数が普及しているのですが、有効性が確立しているわけではないというのも申し上げたとおりです。

その原因の1つは性能が不十分、もう1つはCADの使い方が不適切ということがあると思います。でも、大きいのは性能が不十分ということです。それに対する対策としては、ひたすら性能を上げましょう、今注目の深層学習を使いましょうというのが、有名な米国の会社がやっていることです。色々な技術を組み合わせましょう。Alpha Go（アルファ碁）だって、別に深層学習だけではないですよね、色々なものを組み合わせています。IBMは画像だけではなくて、それ以外のものも組み合わせています。いずれの方法論を採るにしても、データベースが重要であるということは間違いありません。

私の発表は以上で終わるのですが、最後にまとめのスライドを作ってみました。CADの一層の普及のために性能向上というのは必須であって、深層学習は現時点で最も手軽かつ有効な方法かもしれません。問題の見極めというのは必ず必要になると思います。既存の技術との組合せも、もちろん有効でどんどん進めるべきで、これらの研究を組織的かつ計画的に支援することが重要ではないかと思います。

あとはデータベースなのですが、1月に医療ビッグデータ新法が国会に提出されました。個人情報でガチガチに縛るのではなくて、研究目的であればということで例外措置を作るということです。そういう新しい法律の中で、その認証される機関を是非増やせる方向になっていくと、このあたりも充実して良いのではないかと思いました。以上です。

○光石部会長　清水先生どうもありがとうございました。それではまた質疑、応答の時間にしたいと思いますが、いかがでしょうか。

○森委員　おそらく機械学習の仕組みが出てきて、このように使われていると清水先生の発表にあったと思うのですが、今後、進むとなし崩し的にみんなが使い始める可能性があって、WEB上で機械学習のフレームワークは用意するサービスを使ってきて、病院の先生がそれぞれ画像データベースを突っ込んで、もう1つ突っ込んで、あとはもう自動的に、突っ込んでみて、ブラックボックスでもいいから使ってみるという人たちがきっと増えて

くると思うのです。このあたりは少し考えなくてはいけなくて、清水先生はそれに対して警告を鳴らしているのですが、一方、そのまま突き進んでいってしまう可能性が十分にあると思うのです。

○清水委員　その可能性はもちろんゼロではないとは思います。ただブラックボックスのままでは、おそらく性能は伸びないのではないかなというのが1つあります。最近、これはNIPSという学会で報告された論文です。NIPSというと深層学習とかニューラルネットワークの研究がたくさん報告されているところなのですが、深層学習における最適化関数の姿を理論的な面から明らかにした話なのですが、実は3層まではあまりなかったのですが、4層以上に増やすと鞍点が増えるのです。いわゆるsaddle pointと呼ばれているものです。今、最適化の主な方法は勾配法で、1階微分、1階の導関数しか使わないのですが、1階の導関数を使っていてはとても分からない鞍点が増えてくる。それから鞍点の中にも良い鞍点、悪い鞍点があるということです。今、最適化しようとしている関数がだんだん見えてきたのですが、その姿はかなり壮絶な感じで、まともにきちんと解けそうにない。解けそうにないということは性能が出そうにないのではないかという批判があります。ただこれはもちろん問題にもよると思うのです。問題によっては、とてもうまくいく例も出てこないとは断言できないと思います。

だから今、森先生がおっしゃたような、たまたま使ってみて、やってみたら何か知らないけれども、とても性能が出たということはあり得る話だと思います。その出たときにどうするかです。もしユーザーが使って良いと考えればそれを使うということも選択肢の1つとしてはあり得るのだと思います。

先ほどの色々な使い方についての注意事項はあるのですが、使うということもあり得ますが、1つ問題は、何か問題が起こったとき、例えば誤診をしたときに、なぜそれが起こったのか。次にもっと性能を上げようとするときに、どうしたら良いのか。ほかの情報と組み合わせるときにはどうすれば良いのかというあたりがブラックボックスのままですと、ほとんどまた振出しに戻ってしまいます。そのあたりの難しさはどうしても残ると思いますので、理論的な解析を是非、急ぐべきだと思っています。

○森委員　おそらく先生方も9割でもいいから、これで出ればいいではないかという、そういう使い方の人が絶対出てきそうな気がするので、やはりきちんと指針を示したほうが良いと思います。それからおそらく今でも結構WEBベースで画像をアップロードすると適当に機械がやってくれて、分類

してくれるサービスなどがいくつかあることは間違いないので、そういったサービスとどのようにして連携させるか。あるいはどこで仕切り線を書くかなど、そのあたりは是非、議論ができれば良いかなと思います。

○光石部会長　ほかにいかがでしょうか。

○佐藤上席審議役　PMDA 上席審議役の佐藤です。どうもありがとうございました。今日のお二人の先生のお話をお伺いしていると、実際の AI の技術以上にキーとなるのが、先ほど清水先生もお話になった良質なデータを十分に与えられること。これが、非常に何かキーワードとして出てきたのですが、どのぐらいこれが必須というのか、本当にそれがないと駄目なのかどうなのかといったことを、お二人の先生に私見でも結構ですから、教えていただければと思います。それから良質なデータが十分という、結構言うは易しで、ではどの段階で、例えば製品を出す際にどのぐらいのものがと、誰がそこを確認すればよろしいのかという、その 2 つについてもし何かお考えがあれば教えていただきたく、よろしくお願いいたします。

○清水委員　では私からお答えいたします。これは完全に問題依存だと思います。どれだけの数というのは問題依存になってくる話だと思います。前回の松尾先生のお話の中に 1 カテゴリーいくつというような数字が出てきましたが、あれも問題が決まっていればそういう数値を出すことができると思うのですが、どの問題にも当てはまる数がいくつですか？というように質問されると、それはやはり分かりませんというのが、申し訳ないですがお答えになるかと思います。

次に、誰がどのタイミングでということですが、それはやはり薬事の審査の話にもなると思いますので、然るべき数を確保した上で、テストをして数値を評価して、例えば医師単独のときの性能と医師が CAD を使ったときの性能を比較して、同等以上の性能が出ればというようなことで、認可をする方法が 1 つあるのではないかと思います。ここで同等以上と申したのは、先ほど最初のほうのお話でも、人とコンピュータの結果は互いにオーバーラップしないという話がありました。医師が得意な問題とコンピューターが得意な問題というのがあるかもしれませんので、互いに補完し合えば、「同等」でもある程度、臨床的には使えるものになるのではないかと思っています。

○宮野教授　良質なデータを大量にとは本当に言い易い言葉ですが、例えば PubMed が今、2,600 万件ぐらい登録されていますが、そのアブストラクトを今の自然言語処理のツールで、例えば IBM が提供している Watson の API があるのですが、それで 1 件処理するのにコストが 35 円もかかるのです。そうするとたくさんのデータを自然言語に関しても処理するというのは、

結構コストがかかることなのです。そうすると悪書は読まないということが必要になるけれども、悪書も読まなければと、悪書の中にも多少良いところもあるというのを読んでいるのが、私は今の現実だと思うのです。

2,600万件、これ、PudMedですから古いものもあれば、新しいものもありますが、古いものはほとんど利用価値がなく、ランクがわっと下がってきているというのが現実だと思います。例えば薬がターゲットにしている遺伝子、これがどれくらい有効かというのも、治験のフェーズ3が進んでいるというようなことが複数、似たモレキュールであるということであれば、エビデンスが上がってくる。そういったデータは非常に価値が高いのです。だから例えば先ほどの私の話ですと、日本は統一されていませんが、治験情報を統一するなどということは、一遍にデータの量を、ある意味で良質なデータをわっと増やすことになると思うのです。

○佐藤上席審議役　先生、そういうのはやはり人海戦術でやらなければいけないのですか。

○宮野教授　人海戦術が同時に走らないと、駄目というのが今の私の認識している現実です。

○佐藤上席審議役　ありがとうございました。

＜議題1：③医療機器審査部からのプレゼンテーション＞

○光石部会長　まだ質問はあるかもしれませんが、最後にまとめて色々議論する時間もありますので、そのときにまたしていただくということにいたします。3番目の講演に移らせていただきます。次は医療機器審査部から「従来型の医療機器審査基準とアップデート時対応についての情報」ということです。よろしくお願いします。

○医療機器審査第一部審査専門員　PMDAの医療機器審査第一部から医療機器プログラムの審査の考え方について、これは特別に新しい話というわけではないのですが、現状はどのように審査しているかの観点での御説明です。審査の具体的な話に入る前にまず、現在承認を得ている医療機器のプログラムにどのようなものがあるかに関して簡単にまとめたのが、こちらのスライドです。目的性については大きく3点です。診断支援、治療計画、治療支援です。例えば一番上の血管のCTのデータなどを入力して、その中に流れる流圧、流速などがどのようになるかを物理計算するようなソフトウェアであるとか、あるいは放射線治療計画、こういうような角度からこれくらいの線量で打つとどのような線量分布を作るかという、これは物理的な計算をするようなソフトウェア。あるいはバーチャル空間上にものを置いて、どのくらいの長さのワイヤーで治療すれば矯正がうま

くいくかといったようなシミュレータ関係。あるいは既に臨床上で使われている、ある程度分かっているような決まった計算式をうまく応用して、血行動態をシミュレートしてみたり、透析量をシミュレートしてみたりなど、そういうソフトウェアが主になっております。

その中でも先ほど清水先生からの講演にもあったように、CAD と呼ばれるようなものも一部承認を得ている。このような実態になっております。

では具体的な審査の話です。医療機器プログラムの審査というと、ソースコードを見るのですかというような質問をされることもあるのですが、実はこれは誤解で、基本的には医療機器のプログラムをパソコンに入れるなら、パソコンに入れて、インストールされたパソコンが医療機器として役に立つか、適切な動作をするか、そういった観点で見ているということをまず、前提として御理解いただけると非常に助かります。そういった意味で今回、審査におけるポイントとして大きく 2 点まとめました。詳細は様々ありますが、実は去年 3 月に医療機器プログラムの審査に関するガイダンスが出ており、詳細な論点としてはこちらに書いた項目があるのですが、細かく全部説明はできませんので、関連する場所を色付けして、基本的には大きな 2 点で説明いたします。

まず 1 点目です。意図した臨床上の意義が達成できていますか。これが審査の一番大きなポイントになると思います。少しかみ砕いて申し上げると、「このように役に立つ医療機器です」といって流通させたいというならば、「このように役に立つ」ということをきちんと評価できていますか。そんなイメージになるかと思います。

少し具体的な例で御説明いたします。例えばある診断をする行為に対して、通常は 1 年に 1 回ぐらいのスクリーニングをしていて、そこで陽性疑いだった方には次の診断、ここで確定診断をして、陽性確定した方が例えば、副作用重篤ではあるが、そういう治療フェーズに進む。こういう診断のスキームがもともとあるといたします。今回、このスクリーニング用のプログラムとして申請をしたいと考えてくると、審査のときにはどんなことを考えるか。スクリーニング用ですから、恐らく多くの陰性の集団の中から陽性らしい患者を捜してくる、そういうことがきっと達成できているのだろうな、そういうことが評価されていますか、こんな目で臨床的意義とその評価の観点をリンクさせるわけです。

またこういった診断系のプログラムであれば、恐らく感度及び特異度が 100 パーセントというのはさすがに実現できないだろう。ある程度偽陽性、偽陰性が出る。これは仕方ないと思いますが、その偽陰性、偽陽性は許容できるかどうかということを、どのように考えるかということになり

ます。もちろん位置付けはいろいろありますが、例えば、スクリーニングにおいて偽陰性は果たして問題はないか。これは例えば普通に人が診断すれば陽性だと見分けられるのに、プログラムを使って陰性になった、つまり偽陰性になった場合、そのプログラムの罪は非常に重いと考えます。なぜならば1年間、次の診断に行く機会を失わせているわけです。スクリーニング時における偽陰性というのは非常に罪が重いはず。偽陰性が高いとなってしまうと、果たしてスクリーニング用のプログラムとして承認できるのだろうか。そのような位置付けと性能の評価にもリンクしてくるわけです。

一方で確定診断の部分で使うプログラムとして出していきたいのだというように考えるならば、恐らく陽性疑いの患者のデータがいっぱい入ってきて、その中で正しく陰性、陽性が判断できますかということを評価してきましたかと、そういう目で審査をしていきます。もちろん今回、あえて条件として副作用重篤という条件を付けたので、偽陽性、間違った陽性は不要な介入というリスクをどんどん生んでしまいますし、誤った陰性はもちろん誤診ということで問題になる。

このようにそのプログラムがどのような位置付けで使われますかということが、実は審査においては非常に重要なことになってまいります。あとは位置付けに応じて陰性、陽性になりやすい条件等もあれば、情報提供すべき内容というものを、要は使われ方ですが、こういった位置付けで使われますということも判断してきますので、そういったことも審査では確認していくことになると考えます。

今お話したのが、臨床上の有用性をきちんと考えて、それに対して評価できていますかという1つ目の論点でした。

2つ目の論点は、性能評価が適切に行われているか。こちらはどちらかというと、エンジニアリングに近い発想だと思います。プログラムは基本的には何らかのデータをインプットして、処理されてアウトプットが出てくる。このアウトプットが診断なり治療なりに役に立つように使われることになると思います。例えばインプットの情報が1つのケースしかなかった。本当はとてもたくさんのパターンを入力して使われるはずなのに、1例、とても良い結果しか出ないかもしれない情報だけをインプットして非常に良い成績が出ましたと言われても、やはりそれは本当に現場で使える評価になっていますかというところに疑義が生じます。ある程度、実臨床で使われるような状況が想定できるように試験条件というものは組まれていますかというのが重要になってきます。

またアルゴリズム上もともと苦手だというようなデータセットがあれば、

そこに関しても特異的に評価する必要が出てくるかもしれません。そういった試験上の妥当性の評価。アウトプットのほうはもう少し明確で、例えば精度が求められる装置、プログラムだとして、非常に大きな誤差基準でも試験上パスしますというような試験系を組まれてしまっていたら、その結果がパスだからといって、本当に使えますかという議論はどうしても生じてしまいます。この基準というのも妥当ですかということはやはり重要な観点になってまいります。その他、例えばとてもパワーの要るパソコンでしか、あるいはスーパーコンピューターのようなものでしか使えませんというようになってしまった場合には、そういう規定をしなければいけなくなってきますが、本当にその実使用環境と試験環境は整っていますか、一致していますかというところも1つの観点になります。以上のいろいろなことを考えながら、臨床現場に出したときにも発揮できるような条件での性能評価がなされていますかというところが、2つ目のポイントになります。

このようなことを考えながら最後に何を判断するかと申しますと、薬機法、いわゆる薬事法の中で承認拒否事由というものがございまして、この事由に当てはまっていないかどうかを考えています。特に審査の中ではこの①に書かれている効果又は性能を有すると認められないとき、有効性がなさそうとなると、拒否事由に該当して承認ができない。あるいは2つ目、効果に比べて著しく有害な作用があったために使用価値がない、有効性はありそうなのだけれども、それ以上にリスクのほうが大きい、これを使うことによるリスクのほうが明らかに大きいとなってしまったら承認できない。そういうことに当たらないというところを確認しながら承認審査、最後の判断に持っていくことになります。

以上が審査の大きな論点の考え方になります。その他、最後に少しだけ、今後の予想としてですが、AI を用いた機器で想定されるような特徴を書かせていただきました。大きく 2 点挙げさせていただきました。1 点目は正に清水先生のブラックボックスになるかもしれないというところ、機器の特性の把握が困難になる可能性があります。例えばディープランニング、特徴量を自らプログラムが作り出してしまうようなことになってしまった場合に、エラー発生条件と苦手とする部分が、全然メーカーの方でもよく分からないようになってくると、どういう評価系を組んだら適切な評価が果たしてできるのだろうというところは、よくよく考えていかなければいけないなというのが1つございます。

2 つ目、こちらのほうがインパクトがありそうですが、リリース後に臨床現場でどんどんデータを入れていって、育てていく、学習されていく。

そんなようなプログラムを出していきたいとなった場合に、どのようにしてリリース後に変化していくようなものを評価したらよいのだろうか。あるいは規定していったらよいのだろうか。これはかなり難しい問題だと思います。現状の審査では承認時の性能がずっと維持されるという前提での承認が基本になっておりますが、今後こういったものにどのように対応していくか、ここも1つ大きな課題になるかなと考えております。

その他審査以外の論点としても、先ほども出ましたように個人情報保護法の観点もあるので、学習に用いるデータは何でも学習に使ってもいいのか、同意を得る人がいないのか。そんな整理も今後は必要にもなるでしょうし、あるいは市販後学習によって性能が低下して、それを使って誤診した影響だとしたらメーカーが悪いのか、いやいやそれを育てていった現場の使い方が悪かったから、現場に責任があるのかなど、そういう責任問題は付いてくる複雑な問題なのかなと考えております。少し雑多な話になり恐縮でしたが、以上になります。御清聴ありがとうございました。

○光石部会長　ありがとうございました。それでは、ただいまの御講演に対する質疑応答の時間としたいと思います。いかがでしょうか。

○山根委員　医療機器の申請があったときには、従来の医療機器との比較を必ずさせます。プログラムについてもこれまで普及しているようなプログラムだったら、例題を学会が設定して比較するというロジックがあると思うのです。ただ、ここで議論されているような、AI を使った、これまで診断できなかったようなことが診断できるプログラムだという申請が来たときに、比較ということができない場合にどうされるのでしょうか。

○医療機器審査第一部審査専門員　まさに、非常に新しい医療機器を想定されたお話だと思います。そうなってしまうと、おそらく新医療機器の可能性が出てきて、要は臨床試験もきちんとやって効果まで見て、正しく臨床現場に導入できるだけの有用性・有効性が発揮できるかというところを見る、そういったフェーズも十分考えられると思います。

○光石部会長　今までも AI のようなものが入っているものもあるだろうし、これから審査に持ってこられるものも AI と言っていながら違うものもあるかもしれない。ただ、AI と言っても何を指して AI と言えるかというのもあるのですが、例えばディープラーニングではないものも、AI だと言って持ってこられるところもあるかもしれないというのは考えられますよね。

○医療機器審査第一部審査専門員　AI が何かは、よく考えなければいけないと思います。おっしゃるとおりだと思います。

○光石部会長　従来認可されているものを否定するような議論になってしまっては、ま

た困るのではないかという気もするのです。そのあたりをこの専門部会としてどう考えていくかというのは、後で少し議論をしてもいいのではないかという気がしているところです。もし講演内容の質問があればと思いますが、いかがでしょうか。

○鎮西委員　AI というか、機械学習のアルゴリズムが、非常に大きなデータセットを持って育つことを前提とした上で、治験という極めて限られたデータ数のもので評価をするということが、若干イロニカルに聞こえます。ただ、治験という非常にコントロールされたデータ収集という側面はあるのですが、非常にスペクトルの広いデータを基に作った場合のものに対して、例えば狭いデータのものを入れた場合というのは、バイアスの話はほぼなくなってくるわけです。あとは何を見るかですが、エンドポイント的な考え方というのは、かなり変わってくるのではないかという気がするのです。その辺はいかがでしょうか。

○医療機器審査第一部審査専門員　かなり難しい議論だと思います。AI 独特の良さを生かすとなると、単純な臨床試験を組むのが難しいという話は、確かに容易に想像できます。しかし実現性というのも加味して評価を考えていかなければいけませんので、何でもかんでも臨床試験が良いという話ではないというのも十分理解できます。ただ何のエビデンスもなく出していくというのも難しいので、一体どういう折衷案が出せるかというのは、むしろこういうところで議論をして、一緒に勉強させていただきたいというのが正直なところです。

○許委員　私は今、病院をやっています。画像診断は、CT でもかつては 20～30 枚だったものが、桁が 1 つ上がってしまっているわけです。それにものすごく人手を食っていて、院内だけでは処理し切れないような状況になっています。

学生教育から研修医教育を考えてみますと、例えばある画像診断において、我々は教科書で勉強するわけです。アトラスのようなものがあって、それを学習しながら現場である程度の経験を積んで、例えばあるレベルの診断が 95％の正答率で、フォールス・ポジティブ、フォールス・ネガティブ等を全部やって、誤診率がこのレベルに達したら一応専門医とか、そういうイメージで画像診断の専門医を想定するわけです。

そうしますと、例えば画像診断の国家試験的なもの、典型的なもの、これは診断しなければいけないという画像集を 1 万枚くらい作って、それをザッと機械に診断させ、そのフォールス・ポジティブ、フォールス・ネガティブ、正答率、その他全部を評価し、あるレベルを超せばそのプログラムは製造販売承認するという、むしろ臨床的と言うと非常に難しい

のです。ところが実際に我々は、医学をある典型的な画像で勉強しているのです。そういう形の審査が使える部分というのはないでしょうか。

○医療機器審査第一部審査専門員　可能性としてはあり得るのではないかと思います。公的なデータセットと、どれぐらいの基準が達成できていれば臨床上有用と言えるかというのは、何かコンセンサスの取られたものが出てきた場合には、もしかしたらそれをヒントに、それを使ってということもあり得るかもしれません。ただプログラムの開発上、入力情報が分かってしまっていたら、それ用にカスタマイズして作ることもできるので、果たしてそんなことがいいかどうかというのは、評価の中で考えなければいけません。ただ、考え方の方針としてはあり得る話かと思います。

○森委員　先ほどスーパーコンピューターの話も出ていたと思うのです。私は情報基盤センター長をやっているので、スーパーコンピューターを運営しているのですが、おそらく追加学習になってくると、ローカルの臨床器ではできなくて、結局データをどこかに転送して学習させて戻すという作業になって、臨床機器がほとんどコネクティット型の機械に代わってくる可能性があるのです。学習はもうどこかのリモートセンターでやらざるを得なくなってくることは間違いないと思いますから、その部分まで考えておかないといけないかと思います。

○光石部会長　これは宮野先生あるいは清水先生にお伺いしたほうが良いのかもしれないのですが、私がきちんと理解してないのかもしれません。論文に出てくるものというのは、新しいというか、希少疾患などで、通常の医師であれば当然診断できるようなものは、論文には出てこないわけですよね。新しい論文のようなものだけでやって、認識できないといけないものを認識できない、陽性であるものを陽性と言えなかった、そういうものを見落とさないかということが、多少心配になってくるのです。そのあたりはどうでしょうか。

○宮野教授　論文やリポートで出ているものが、どの程度のエビデンスになるかということがあるかと思います。例えば『The New England Journal of Medicine』などには、症例が少なくても載ります。そうすると結構皆さん、すぐに自分の所で使えないかと判断します。一方でそれがバイオロジー系のところに出ていて、例えばがんの細胞株を使ってこうだったというような実験で、この化合物がこの遺伝子の発現を抑えてやると増殖を止めたなどというのは、ある意味ではエビデンスが低いのです。ただ、がんに関してですが、私たちは、がんの多様性と複雑さというオブスタクルの前では、どんな知識でも情報でも、使えるものは使いたいなと思います。「藁をもすがる」という言葉がありますけれども、そんな感じ

で見ています。

今回の先生の横からの回答にしかなっていませんが、普通の学会のガイドラインで、診断が付くフローチャートのようなものがありますね。それで診断が付かないもの、あるいは診断が付いたものでも、随分ひっくり返るということを、この2年間、医科学研究所の現場で見ています。そうしますとゲノムを見るようになる前と、見るようになってから後というのは、随分と認識が変わったところです。血液の場合は時系列でどんどん変化して、ある薬で完全に寛解し、標的にしたクローンが完全にゼロで、デジタクルPCRで見てもワンドロップレットも出てこないことがあります。それで普通は寛解して良かったということですけれども、それが半年後ぐらい経ってまた同じものが、ドロップレットが出てくるということが普通にあるのです。

それを論文に発表できるかというと、そういうものはあまり論文にもならないことなのです。そうすると、こういうツールを使いながら、知識としてソフトウェアというか、データベースなどに貯まっていくものしかないのではないかと。だから文献などは、あくまでもそういうものに到達していく1つのガイドラインというか、掴まっていく手すりのような感じが私はしています。

＜議題2：検討の方針と今後のスケジュールについて＞

○光石部会長　分かりました。もうちょっと色々議論したいのですが、時間も限られていますので、改めて別の機会にさせていただきたいと思います。

それで、前回までのものではあるのですが、資料5に今まで出てきた論点整理ということで、質問事項や回答等をまとめています。これを見ますと、変容し続けるシステムということで、後で学習させるものかどうかとか、データベースの利用・信頼性、ディープラーニングの特徴、臨床応用、性能保障と承認の在り方、責任の所在、倫理、そして課題として挙げられた点、その他といったものに、ある意味まとめられます。

それから、資料1を見ていただきたいと思います。ただこれだけをここでまとめたのでは、やはり足りないところもあるのではないかということで、技術の俯瞰と課題抽出とか、これは資料5になるのかもしれないのですが、新要素、規制、ベンチマークと承認の在り方、倫理・責任といったまとめ方のほうが、もう少しいいかもしれないと。報告書と言うにはちょっと早いかもしれないのですが、こういうところで議論をしていってはどうか、まとめていってはどうかというたたき台を作っているところです。残りの時間は、これについて御意見をいただければと思いま

す。いかがでしょうか。それと、先ほどの既存のものをこの中にどう入れ込んでいくかというあたりもですね。

○鎮西委員　資料5の2ページの5.の一番上に、私の発言が取り上げてあるのですが、私がこのときに質問で問いたかったのは、どちらかと言うと市販後安全対策という観点もあるという意味だったのです。そういう点も考えておく必要があるのではないかという具合に御理解いただければと思います。

○光石部会長　これは大雑把にということですので。

○鎮西委員　あと、もう1つよろしいですか。厚労省のAI懇談会との切分けというか。同じことをやってもしようがないような気がするのです。これはどうするのですか。

○大江副部会長　AI懇談会のほうは来週に最終報告を出します。こちらのほうがスケジュールとしてはだいぶ後ですから、内容がぶつからないようにできるのではないかと思っています。

○鎮西委員　では、AI懇談会で議論されたことを確認するという感じで大丈夫だと。

○光石部会長　ほかにいかがでしょうか。

○石塚委員　宮野先生に確認します。今回、血液腫瘍のWatsonとマニュアルの解釈の話で、1度の違いが出てきたのですが、ほかの腫瘍の例でも大体同じような感じなのですか。

○宮野教授　症例数が少ないので御報告していませんが、大腸がんなどの消化器がんとか、先ほどの虫垂由来のものなどは、全く診断が付かないものにどう付けるかということなので、うちはスタティクスを出せる段階ではありません。血液の場合は比較的患者が多いということと、全ゲノムシーケンスの解析はほとんどやっていませんけれども、エクソームの解析をやるという格好で、費用的にも何とか流せていけているのです。それと、実際に最初の病理診断からひっくり返って、患者が良い方向に向かったという実績があるので、そちらのほうの数が多くなっているということです。ですから固形がんのほうは統計を出せる状態ではありません。ある1、2例について、有効であったかどうかという御報告にならざるを得ませんでした。

○光石部会長　特にこのあたりを重点的にというものは、大江先生から何かありますか。

○大江副部会長　評価手法をどう決めていくのか。やはり有効性の評価、あるいはリスクの評価に関する部分というのを、今の第三次AIブームで取り上げられるような技術が、医療現場で実際に応用されていくときにどうなるのかという視点は、少し力点を置いて書く必要があるかと思っています。

○光石部会長　ほかの委員の方、いかがでしょうか。

○佐久間副審査センター長　今日のお話は、やはりデータの妥当性の話とともに、問題依存性という話がいくつかありました。宮野先生のお話では非常に多様性が高く、それがゆえに非常にまれなものでも、情報としてはすごく有用だというものがある一方で、それがすべてのがんかどうかまではよく分からないかもしれないと。

それから、先ほどの清水先生のお話の中では、得意な問題、不得意な問題というのが、実はまだアルゴリズム的にも理論的にも分かってないので、そこがまだクリアに言い切れていない部分があるかもしれないと。そのあたりが今、AI ブームといったときに、何でも使えそうだというところに対して、どういうところに落とし穴があるのかといったことを、きちんとまとめられても良いのではないかと思ったのですが、そのあたりはどうでしょうか。

○清水委員　ある意味、本質的な問題だと思います。希少がんの話のように少数例がきちんとあれば分かるという例もあります。ロングテールの話でよく説明されるのですが、世の中のほとんどの問題はロングテールなので、例をたくさんの種類集めてることが、まずは大事だと、そういう話があります。その一方で、メジャーなところもきちんと押さえておかないといけないという場合には、データ数も必要だと。

両方の問題を一遍に解こうとすると、すごい数になってしまいます。ロングテールの一つ一つの問題を大事にしようという場合は、それが得意な AI もありますから、その技術に必要なデータベースはこういう方針で作っていけばいいという話があります。そうではなく、全体をどうしようという場合、細かい分類はできないけれども、とにかく拾い上げができるようなシステムを考えようというようになると、できるだけたくさんということになってきます。そのあたりはきちんと問題が整理できると、すごく良い話になってくるのではないかと思います。

○大江副部会長　今の話は多分、大規模なデータセットから学習して開発していくようなシステムには、今後非常に重要な課題というか、問題になっていくという気が、私も佐久間先生のお話などを聞いていて思いました。分かりやすく言いますと、例えば多数の肺がんの画像だけで十分学習した胸部X線写真診断支援システムというのは、結核は見落とすかもしれないわけですよね。そうすると、この機器はこういう問題のデータセットで学習したシステムであるというスコープ、つまり適用可能な領域を、いかに使う側に正しく提示するかということも大事になってくるのではないかと、今聞いていて思いました。そういうスコープの提示の仕方が、どういうようにあるべきかというのも議論しないといけないような、そんなこと

を感じました。これはコメントです。

○光石部会長　これはあくまでも案ですので、この辺のまとめ方についてはまた見ていただいて、もう少しこういうようにしたほうがいいのではないかということがありましたら、是非事務局にお伝えいただければと思います。次回、第3回専門部会では更にAIの応用を俯瞰するということで、[redacted]予定にしております。[redacted]そういうことで、本日の議論は以上とさせていただければと思います。では、事務局から連絡事項をお願いいたします。

＜議題3：その他＞

○事務局(江原)　次回の専門部会は1か月後、4月20日木曜日の午前10時から予定しております。どうぞよろしくお願いいたします。場所はPMDAの予定ですが、また御連絡・御案内を差し上げます。

＜閉会＞

○光石部会長　それでは、本日の専門部会はここまでとさせていただければと思います。どうもありがとうございました。