

医療 AI と医療データ 2 次利用に関する話題

森 健策 (名古屋大学 大学院情報学研究科 教授／名古屋大学 情報基盤センター センター長／国立情報学研究所 医療ビッグデータ研究センター センター長)

Topics on Medical AI and Secondary Use of Medical Data

Kensaku Mori

*Professor, Graduate School of Informatics, Nagoya University / Director, Information Technology Center, Nagoya University /
Director, Research Center for Medical Big Data, National Institute of Informatics*

【要旨】 近年の人工知能 (Artificial Intelligence) 技術の発展は目覚ましく、社会の様々な分野において AI が利用されるに至っている。画像を分類する分類型 AI、画像から物体を検出する画像検出 AI、対象物体の領域を抽出する画像セグメンテーション AI など多数の画像認識が登場している。医療分野も例外ではなく、医療の様々な分野で AI の利活用が進みつつある。例えば、内視鏡画像診断を支援する AI、X 線 CT 画像といった放射線画像の診断を支援する AI、血液検査や心電図などの検査データなどの診断を支援する AI など、様々な医療 AI が登場している。医療分野での AI を考えるとその AI をトレーニングするためのデータは必然的に臨床の場で取得されたデータである。臨床の場で得られたデータを 2 次利用して AI の学習処理を行うことになる。そこで本稿においては、医療 AI、特に医療画像 AI に焦点を当てながら、AI の研究開発と医療データの 2 次利用について議論したい。

【キーワード】 医療 AI 医用画像 AI 医療データ ビッグデータ 2 次利用

【Abstract】 Artificial Intelligence (AI) technology has made remarkable progress, leading to the use of AI in various sectors of our society. Numerous image recognition techniques have emerged, such as classification-type AI for categorizing images, image detection AI for identifying desired objects within images, or image segmentation AI for extracting specific object regions from an image. These technologies are now used in many societal scenarios. The medical field is no exception, with an increasing adoption of AI. For instance, there are AIs designed to assist in endoscopic image diagnosis, support in radiological image interpretations such as X-ray and CT scans, aid in the diagnosis based on blood tests and electrocardiograms, and analyze electronic medical records. When considering AI in the medical field, it becomes inevitable to use data obtained in clinical settings for training these AIs. This implies the secondary use of clinical data for AI training processes. In this paper, we aim to discuss the research and development of medical AI, focusing particularly on medical imaging AI, and the secondary use of medical data.

【Keywords】 Medical AI Medical Image AI Medical data Big data Secondary use

1. はじめに

近年の人工知能 (Artificial Intelligence) 技術の発展は目覚ましく、社会の様々な分野において AI が利用されるに至っている。画像を分類する分類型 AI、画像から所望とする物体を検出する画像検出 AI、画像中から対象とする物体の領域を抽出する画

像セグメンテーション AI など多数の画像認識が登場し、社会の数多くの場面で利用されている。スマートフォンで実装されている対象物体切り出しは画像セグメンテーション AI の好例であろう。更に最近では、様々な画像やテキスト情報などを AI によって生成する生成型 AI についても注目を大きく得ている。ChatGPT は世の中に大きな影響を及ぼし¹⁾、世界的な枠組み作りなどが G7 なども含めて議論が

進んでいるところである。

医療分野も例外ではなく、医療の様々な分野で AI の利活用が進みつつある。例えば、内視鏡画像診断を支援する AI、X 線 CT 画像といった放射線画像の診断を支援する AI、血液検査や心電図などの検査データなどの診断を支援する AI、電子カルテの分析を行う AI など、様々な医療 AI が登場している。大腸内視鏡診断支援 AI²⁾ (図 1) は、複数社が医薬品、医療機器等の品質、有効性及び安全性の確保等に関する法律（薬機法）に基づく医療機器としての承認を受け、臨床現場に向けた製造販売が行われている。

AI はコンピュータによって人の知的な活動を模倣することを目的として研究が進められてきた。そのコンセプトは 1956 年にアメリカで開催されたダートマス会議までさかのぼる。この会議では、コンピュータ科学分野の研究者らが集まりコンピュータによる知能の実現について議論を行った³⁾。Artificial Intelligence (AI) の語句が初めて使われたのもこの会議であるといわれている。ニューラルネットワークを用いて画像などのパターン情報を認識するパーセプトロンが開発されたのもこの頃である⁴⁾。ニューラルネットワークとは、脳の神経細胞のネットワークを計算的に模したものであり、入力データとそれに

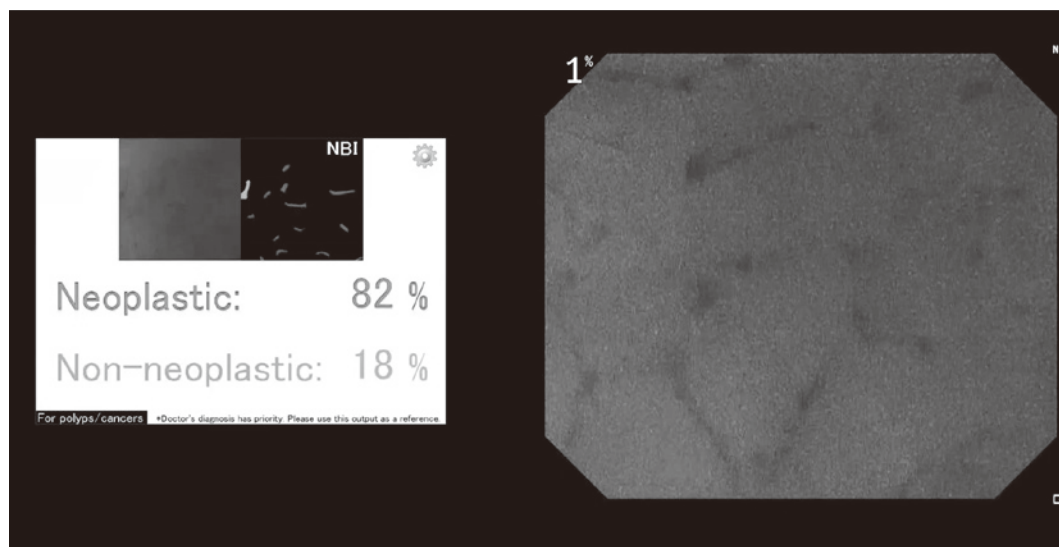
対応する出力データの組を大量に与えることで、入力を与えると所望とするデータが得られるようになるものである。ネットワークの複雑さは、入力に対してどのような複雑なデータが得られるかを表している。様々なニューラルネットワークの構造が提案され、画像認識、音声認識、画像生成、テキスト生成など 1950 年代では想像もつかなかった複雑なタスクを実行することができている。

AI、特にニューラルネットワークを学習させるには数多くのデータが必要とされる。テキスト生成などに用いる GPT (Generative Pretrained Transformer) モデルは、その内部に大量のパラメータがあり、これを決定するために膨大な量のデータが必要とされる。いかにして大量の学習データを用意できるかが AI の精度を決めるといっても過言ではない。

医療分野での AI を考えると、その AI をトレーニングするためのデータは必然的に臨床の場で取得されたデータを利用することになる。臨床の場で得られたデータを 2 次利用して AI の学習処理が行われる。2 次利用が促進されることによって初めて医療分野での AI 開発が進むといえよう。しかしながら、医療データの 2 次利用を行うに当たっては様々な課題があることも事実である。

そこで本稿においては、医療 AI、特に医療画像

図 1 大腸内視鏡画像診断支援システムの例。大腸ポリープ表面画像に関してコンピュータが診断を行った例。左側に腫瘍性 (neoplastic) から非腫瘍性 (non-neoplastic) であるかの判断結果が表示される



AIに焦点を当てながら、AIの研究開発と医療データの2次利用について議論したい。

2. 医用画像 AI 開発におけるデータ活用 の問題点

医用画像 AI 開発分野における利活用の問題点を整理するとすれば、データ収集における問題、データ利活用における問題、AI 開発における2次利用に関する問題点、AI が出力するデータに関する問題点などがある。これらについて議論を進めたい。

2.1. データ収集における問題点

2.1.1. 医療データの1次利用と2次利用

自明なことかもしれないが、ここでは医療データの1次利用、2次利用について簡単に整理をしておきたい。1次利用とは臨床の場で取得される検査データ、画像データなどを診断治療の目的で直接的に利用することである。データの利用は、患者本人の診断治療目的で使われるものであって、それ以外の目的で利用されることはない。2次利用とは、診断治療の目的で取得されたデータを直接的な診断治療目的（患者診療）以外に利用することである。例

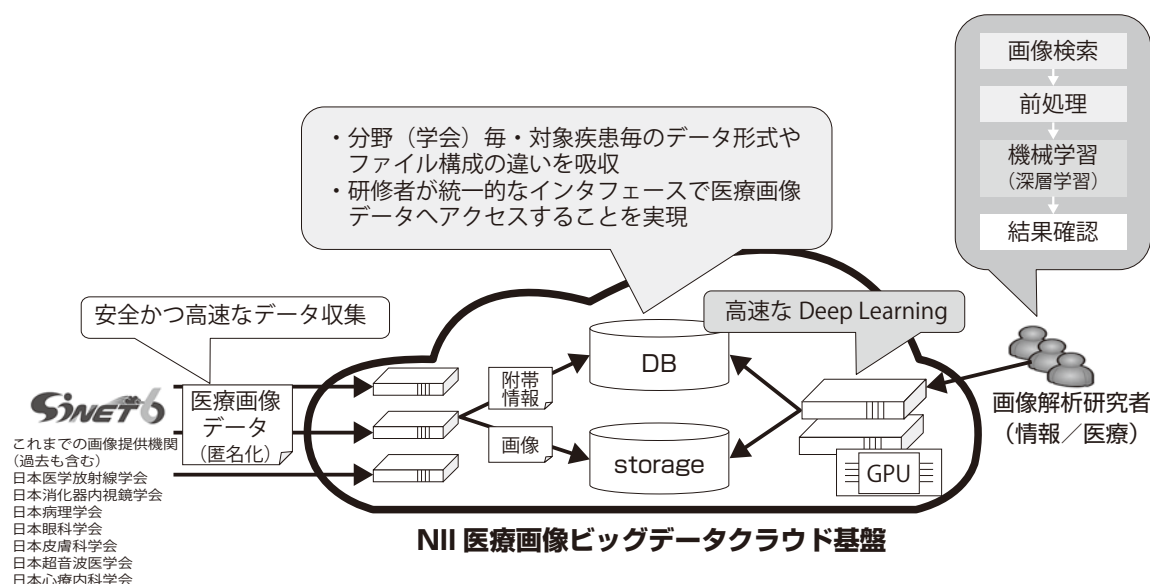
えば、国立情報学研究所医療ビッグデータ研究センターでは、臨床系の学会と協力し、国内の臨床機関で撮影されたCT画像等の医用画像を収集し、医用画像AIを開発可能なクラウド基盤を整えている。大規模に医用画像データを格納することのできるクラウド基盤が用意され、医用画像データの2次利用による医用画像診断支援AIの開発が進められている（図2）。

なお、2次利用で得られたデータ（この場合には何らかの形でデータが加工される）をさらに次の目的に利用する3次利用、4次利用の定義をすることも可能であるが本稿では、2次利用までを考えた

2.1.2. データ収集における問題

医療向けAIを開発する場合には、日常の診療業務で得られるデータを利用することになる。AIの開発に必要とされるデータを取得するために別途検査を行うといったことは行われない。したがって一般的には診療業務で得られるデータを2次利用することが一般的であろう。このようなデータを収集して、AIの研究開発に用いる場合には、個人情報保護法、次世代医療基盤法、人を対象とする生命科学・医学系研究に関する倫理指針^{3,4)}、などといった関連

図2 国立情報学研究所医療ビッグデータ研究センターが運用する医療ビッグデータ解析基盤の例。国内の病院で撮影される医用画像を収集し、国内大学の研究者に対して医用画像診断支援AI開発基盤を提供している。



する法律や公的なガイドラインに準拠して収集し、AI 開発に利用可能とする必要がある。データ収集の対象となる被験者からの同意取得も重要である。医師が患者ごとにデータ利用に関しての同意を得るオプトインと呼ばれる手続きでの同意取得、あるいは、データ利用を AI 開発のためのデータ開発に用いることに同意しない場合には、別途申し出るようにするオプトアウトなどがある。オプトインの場合にはデータの利用目的などの説明書が直接手渡され説明が行われる。オプトアウトの場合には、データ利用に関しての掲示が病院等において行われ、データ利用に同意しない場合には申し出て欲しい旨の記載がなされる。

「はじめに」で述べたように、高精度な医療支援用の AI を開発するには大量のデータが必要とされる。オプトインで個別の同意を得ることが原則必要であり、データ収集するには極めて長い時間を要す。一方、過去に取得されたデータ（例えば CT 画像）などを AI 開発に利用したいということもあろう。この場合には郵送等によって患者同意を取得する必要があるといった問題点がある。オプトアウトを利用すれば、病院等での掲示やホームページでの掲載によって、過去の取得されたデータを 2 次利用できるようになる。学術機関（大学等）学術研究に用いるのであれば問題はないが、収集されたデータを企業が商品開発に利用する場合には注意が必要とされる。

継続的に収集されたデータのデータベースを利用する場合も注意が必要とされる。医療データを収集する場合には、ある目的に沿ってデータを収集することが行われる。このようなデータは、他の目的に利用することができる。例えば、肺がん診断支援 AI の開発を目的として収集されたデータベースがあったとしよう。このデータベースには大量の胸部の CT 画像が格納されている。そのために、例えば、肋骨骨折自動検出 AI にも利用することは当然できる。肺がん診断支援 AI 開発を目的としたデータベースを肋骨骨折自動検出 AI の開発に利用すると目的外利用となる可能性がある。利用目的の追加を行うか、データベース構築における検討の際に、格納されるデータの利用目的の十分な検討も必要であろう。

次世代医療基盤法は、データ収集とその利用に関する問題を解決するために作られた法律でもある。しかしながら、次世代医療基盤法に基づく認定匿名加工医療情報作成事業者は 3 事業者にとどまっており、今後の利活用の促進が望まれる。

2.1.3. データオーナーシップの問題

医用画像の 2 次利用において、そのデータオーナーシップも大きな問題点となりうる。例えば、病理画像検査で用いられる画像について考える。検体は患者から取得されたものであるが、それを顕微鏡で観察できるように切片化、染色処理といった標本作成処理を行い、病理標本スライドスキャナにより画像化（WSI; Whole Slide Imaging と呼ばれる）するのは病院である。AI 開発においては、学習に利用するデータの選択も重要である。大量に病院内のデータベースに格納されている画像から AI 開発に適切な画像セットを選択しなければならない。この選択には相当の専門的知識が求められる。一般的には医師が画像選択に従事することになる。また画像検査にかかる費用については、患者自身の自己負担や公的保険による費用負担がなされるが、そこには公的な資金も含まれる。医用画像データを 2 次利用する場合には、先述の倫理ガイドラインや個人情報保護の観点のみならず、数多くのステークホルダーが存在する。AI 開発においてはこのような画像を匿名化して利用することになるが、この匿名化された画像データの利用権限や帰属についてどう整理するかについて確認をすることが重要であろう。

2.1.4. アノテーションデータの著作権

医療支援 AI を構築する上では、元となるデータに加えてそのデータに対するアノテーションデータ（注釈データ）が必要となる。例えば、肺結節影 CT 画像のデータベースに対して、どこに肺結節が存在するのか、画像上での輪郭線などのアノテーションデータが付与され、このアノテーションデータに基づき AI の学習が行われる。このアノテーションデータはデータベース中の画像に対して付与されたものであるが、データ作成の主体はアノテーションを行う医師である。このようなアノテーションデータの著作権はそれを作成した医師にあると考えられるが、アノテーションデータと元となる画像データ

それぞれのオーナーシップの関係についても整理しておくことが必要であろう。

2.2. データ利活用に関する問題点

2.2.1. 第3者提供

医用画像データの2次利用においては、データ収集を行った研究者によって利用がなされ、画像診断支援AIなどの開発に役立てられる。このような利用は当初の目的に合致しており大きな問題は発生しない。しかしながら、研究開発の進展により、データ収集を行った研究者からさらに第3者へとデータが提供される場合もあり得る。当初は学術機関において、データ2次利用の形でデータ収集がなされてAI開発に利用されていたが、研究の進展において企業との共同研究へと発展し商品開発へとつながるような場合、第3者である企業へとデータ提供がされる可能性がある。このような第3者利用については、個人情報保護法によってさまざまな制限がなされている。倫理指針やそのガイドダンス、あるいは、諸ガイドラインとの整合性なども考慮にいれて判断する必要がある。

2.2.2. 学術利用と商用利用

医療分野で用いるデータは個人情報保護法の対象となるデータも多い。このような個人情報を利用する場合、個人情報保護法に則った対応が求められている。個人情報保護法においては、基本的には患者個人同意が必要としているが、公衆衛生の向上、学術研究目的を対象とした例外規定も定められている。医用画像データの2次利用を考える場合、学術目的での利用なのか、商用目的での利用なのかに応じて、対応も異なる。医用画像データを2次利用して画像診断支援AIを実現するには個人情報保護法は頻繁に改定されており、同法ならびに当該法律に基づいたガイドライン、あるいは、公的なQ&Aに十分に注意をする必要がある。

2.3. 生成系AIが与えるインパクトと医療データ2次利用

ChatGPTに代表される生成系AIは世界中に大きな影響を与えた。ChatGPTといった生成系AIの内部では、大規模基盤モデルを利用している。大規模

基盤モデルを利用目的に応じた追加的な学習(fine tuning)を実施することで所望とする性能を実現している。この大規模基盤モデルには膨大な数のパラメータがあり、このパラメータの学習を行うには膨大な量のデータが必要とされる。医療分野でも利用可能な大規模基盤モデルを作るには、一般的なデータに加えて、医療分野におけるデータも大量に用意し、それを学習する処理が必要となる。この大規模基盤モデルの学習に医療データを利用する場合、2.1.で示したような問題点がより顕在化する。医療データの2次利用として大規模基盤モデルを学習に用いるためには、様々なハードルを越える必要がある。もちろん、一般的なデータを用いて大規模基盤モデルを学習し、それを大規模基盤モデルの学習で利用することに同意を得た少数の医療データを用いてfine tuningするやり方もあるであろう。いずれにしても、これまでとは比較にならないほどデータに貪欲な手法が実現され、その有効性が示されており、医療データの2次利用においても十分に注意する必要があると考える。

2.4. 学習モデルデータのオーナーシップ

大量の医用画像を2次利用することによってAIの学習を進めた場合、そのAIの学習モデル(具体的にはパラメータの大規模集合)のオーナーシップについても整理しておいた方がよい。一義的には学習処理を行った開発者、あるいは、その所属組織にあるであろう。一方、2.1.3., ならびに、2.1.4.に記したことを考えると問題は複雑になるかもしれない。加えて、学習済みのモデルは個人情報に当たるとの議論も必要であろう。学習済みモデルから、学習に利用されたデータそのものを完全に復元することは難しい。しかしながら、学習済みモデルから、学習に使われたデータを復元する試み(model inversion attack)も行われているのも事実である。

2.5. AIが生成するデータのオーナーシップ

2.1.3., ならびに、2.1.4.においては、2次利用するための医療データ収集についての問題点について述べた。一方、医療分野AIが出力するテキスト、あるいは、画像のオーナーシップはどのように整理

されるのかも検討する必要がある。医用画像とそれに付随する所見文の 2 次利用によって学習させた AI を用いて、別途入力される CT 画像などの所見文を自動的に生成した場合、生成された文章のオーナーシップはどうなるのであろうか？ まだ実現されていない技術ではあるものの早急に整理検討する段階にあると考える。

3. むすび

本稿においては、医療 AI、特に医療画像 AI に焦点を当てながら、AI の研究開発と医療データの 2 次利用について議論を進めた。医療 AI を実現する上では医療データの 2 次利用が欠かせない。一方、このような 2 次利用において様々な制約があるのも事実である。関連法令、ガイドラインを遵守しながら、より良い医療支援 AI の開発のために医療データの 2 次利用が促進されることを期待したい。なお、本稿で示す規制等は毎年改正が行われている。特に AI の急速な進展に伴い、改正も頻繁に行われるようになった。最新の情報を収集することが望まれる。

【謝辞】 本稿で記した内容は JST CREST 等における医療データを 2 次利用する研究課題で得られた知見に基づいている。

注・参考文献

- 1 Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, Xia Hu (2023) Harnessing the Power of LLMs in Practice: A Survey on Chat GPT and Beyond, arXiv: 2304.13712 (doi:<https://doi.org/10.48550/arXiv.2304.13712>).
- 2 Masashi Misawa, Shin-ei Kudo, Yuichi Mori, Tomonari Cho, Shinichi Kataoka, Akihiro Yamauchi, Yushi Ogawa, Yasuharu Maeda, Kenichi Takeda, Katsuro Ichimasa, Hiroki Nakamura, Yusuke Yagawa, Naoya Toyoshima, Noriyuki Ogata, Toyoki Kudo, Tomokazu Hisayuki, Takemasa Hayashi, Kunihiko Wakamura, Toshiyuki Baba, Fumio Ishida, Hayato Ito, Roth Holger, Kensaku Mori (2018) "Artificial Intelligence-Assisted Polyp Detection for Colonoscopy: Initial Experience," *Gastroenterology*, Vol. 154, No. 8, pp.2027-2029 (First online: 11 April 2018) (doi: [0.1053/j.gastro.2018.04.003](https://doi.org/10.1053/j.gastro.2018.04.003)).
- 3 J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon (2006) "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955", *AI Mag*, vol. 27, no. 4, p. 12.
- 4 Frank Rosenblatt (1958) "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychol Rev.*, 65(6), pp. 386-408 (doi:[10.1037/h0042519](https://doi.org/10.1037/h0042519)).
- 5 Matt Fredrikson, Somesh Jha, and Thomas Ristenpart (2015) "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures," In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS '15) . Association for Computing Machinery, New York, NY, USA, 1322-1333. (<https://doi.org/10.1145/2810103.2813677>).
- 6 文部科学省, 厚生労働省, 経済産業省 (2023) "人を対象とする生命科学・医学系研究に関する倫理指針," 厚生労働省ホームページ (<https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/hokabunya/kenkyujigyou/i-kenkyu/index.html>).
- 7 文部科学省, 厚生労働省, 経済産業省 (2023) "人を対象とする生命科学・医学系研究に関する倫理指針ガイダンス," 厚生労働省ホームページ (<https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/hokabunya/kenkyujigyou/i-kenkyu/index.html>).