

4. 医療AI研究開発のためのデータ収集・活用の現状と 将来展望

2) 医療AIの研究開発を支える 医療データ基盤構築への取り組み

高木 優介 医療AI推進機構(株)

医療AIの研究開発において、アルゴリズムの高度化が重要であることは言うまでもない。しかし、実臨床で有用なAIを開発し、医療現場へ実装するためには、アルゴリズムと同等、あるいはそれ以上に、学習・評価に用いる医療データの質と量が重要となる。

医療データは個人情報を含む機微性の高い情報であり、研究開発に利用するためには、個人情報保護法や次世代医療基盤法をはじめとする関連法令を踏まえた適切な取り扱いが求められる。そのため医療データは、医療機関から安全に取得し、匿名加工し、研究開発に利用可能な形に構造化した上で、将来的な薬事評価や社会実装まで見据えて管理する仕組みが不可欠である。

弊社は、医療機関から医療データを収集し、医療AI開発企業や製薬企業などに提供するデータ事業を中核に据え、日本におけるsoftware as a medical device (SaMD)の薬事対応支援事業、そして、匿名加工技術をはじめとする医療AIの研究開発を一体的に展開している。これらは独立した取り組みではなく、医療AIの研究開発から社会実装までを支える一連の基盤として、相互に接続している。本稿では、弊社が強みとする放射線画像および読影レポートを中心に、医療AI研究開発を支えるデータ収集、PACS連携、匿名加工、構造化、データベース化、そして、薬事対応を見据えた取り組みについて紹介する。

医療AI研究開発における データ基盤の課題

医療AIの研究開発では、画像データに加え、読影レポート、各種検査結果など、複数種類の情報を組み合わせて利用することが増えている。また、モダリティや検査種別、撮像・測定条件、装置、施設ごとの運用、患者背景なども多様であり、限られた環境で構築されたデータセットだけでは、実臨床における性能や汎化性を十分に検証することが難しい。とりわけ画像診断AIにおいては、画像単体から得られる情報には限界があり、病変の有無、部位、診断名、鑑別診断、経時変化、臨床的意義などを把握するためには、読影レポートや関連する臨床情報との紐付けが重要となる。

一方で、医療機関に蓄積されているデータが、AI研究開発にそのまま利用できる形で保存されているとは限らない。PACSには膨大な画像が保存されているが、研究目的に応じて対象症例を抽出し、必要な画像シリーズを取得し、レポートや検査値と対応づける作業には大きな手間がかかる。また、施設ごとにPACSやレポートシステムの運用が異なり、データの形式や項目名、記載方法にも差がある。

さらに、データを医療機関外で研究開発に利用する場合には、個人識別リスクへの対応が不可欠である。DICOMタグには患者ID、氏名、撮影日などが

含まれるが、画像そのものに患者名やIDなどの文字列が焼き込まれていることもあり、頭部CTやMRIでは顔貌に由来する識別可能性も考慮する必要がある。読影レポートなどの自然言語文にも、氏名、日付、施設名などが自由記載として含まれることがある。したがって、医療AI研究開発のためには、データ取得、匿名加工、構造化、品質管理を一体で設計する必要がある。

放射線画像・読影レポート を中心としたデータ収集

弊社では、放射線画像および読影レポートを中心とした医療データの収集・利活用に取り組んでいる。放射線画像は医療AIの研究開発において最も活用が進んでいる領域の一つであるが、AI開発に適したデータセットを構築するためには、単に画像を大量に集めるだけでは不十分である。どの疾患・タスクを対象とするのか、どのモダリティや撮像条件を含めるのか、読影レポートや検査結果とどう紐付けるのかを、研究開発の目的に応じて設計する必要がある。

特に読影レポートは、放射線科医の専門的判断が記録された重要な情報源である。画像診断AIの開発では、読影レポートから所見、診断名、病変部位、経時変化などを抽出し、画像データと対応づけることで、疾患検出、分類、重症度評価、予後予測など、多様なタスクへの活用が期待できる。一方で、レポー