2. ディープラーニング研究の臨床評価と薬機法審査の実際

三澤 将史 昭和大学横浜市北部病院消化器センター

近年，畳み込みニューラルネットワーク（convolutional neural network：CNN）に代表されるディープラーニングは，画像認識領域のベンチマークテストであるImageNet Large Scale Visual Recognition Challengeにおいて，ヒトの誤答率を下回ることが報告されている。これをもってして，世間一般では，ディープラーニングはヒトの画像認識能力を超えていると認識されていることが多い。しかしながら，これは人工知能（artificial intelligence：AI）に対して，あまりに楽観的であると言わざるを得ない。AIの技術力では，世界最高水準のCNNを搭載しているであろうGoogleフォトであっても，ヒトをゴリラと誤認してしまうことがある（https://twitter.com/jackyalcine/status/615329515909156865）。ヒトであれば，子供であっても瞬時に区別できることが，世界最高水準のAIでもできない。この事例は，学習画像の偏り（バイアス）によって，汎化性能が失われた結果で，リアルワールドでユーザーの報告から初めて発覚した欠陥である。回りくどくなったが，医療においては，このような誤認・誤診は避けなければならない。上市前にリアルワールド，もしくはそれにきわめて近い環境でAIの性能を評価しておくことが必要である。本稿では，主にディープラーニングの臨床における評価方法と，われわれが開発に携わり，薬機法承認を取得した内視鏡診断支援AIの薬機法承認審査プロセスについて紹介する。読者のAI医療機器開発における何らかの一助となることを期待する。

学習データセット構築における注意点

ディープラーニングに代表されるAIは，なぜ臨床（*in vivo*）で評価する必要があるのか，机上実験（*ex vivo*）ではだめなのか？　われわれは，*in vivo*での評価が真の汎化性能を図る唯一の方法であると考えている。*ex vivo*の評価方法と解釈における注意点は非常に重要な事項であるため，本特集の「データセットの作成と学習方法のノウハウ，学習後の評価方法」（49～53ページ）と重複するが，再度ディープラーニングの学習におけるデータセットの構築のポイントについて，医師の立場から説明する。

まず，開発の初期においては，ほとんどの場合，エンジニアは低クオリティデータ（内視鏡画像で言えば，ブレた画像，ノイズが乗った画像，残渣が残っていて評価困難な画像など：**図1**）を除外して，高クオリティの「きれいな」データを可能なかぎり多く臨床医に要求することが多い。これらの画像は，①トレーニングデータ（ディープラーニングの重みを更新していくことに使用），②バリデーションデータ（学習の途中段階において汎化性能を確認し，無数のハイパーパラメータの調整に使用），③テストデータ（学習に使用していない外部データとして，AIの性能を評価するために使用）の3つのデータセットに分ける。一般的には，③テストデータに対する性能が良

〈0913-8919/19/￥300/ 論文 /JCOPY〉