巻頭言　機械学習の「正解」教師データに悩む

三島　和夫
秋田大学大学院医学系研究科精神科学講座

精神医学，精神医療のさまざまな分野でウェアラブルデバイスやクラウド解析を活用して診断や治療を行う取り組みが進められている。とりわけ睡眠医学は以前から睡眠ポリグラフ検査，簡易型脳波計，アクチグラフ，経鼻的持続陽圧呼吸療法（CPAP）などの医療機器からのデジタルデータを解析して診療に還元するデータフローが確立されているため，IT や IoT に親和性が高い。

デジタル技術を活用して診断や治療を支援するプログラム医療機器も医療現場に実装され始めている。医薬品医療機器等法（薬機法）において単体のソフトウェアも医療機器として認められており，不眠障害や睡眠関連呼吸障害への応用が進んでいる。2023 年の 2 月には，禁煙治療用アプリ，高血圧治療補助アプリに続いて不眠障害の認知行動療法の補助アプリ（サスメド株式会社）が国内では 3 番目のプログラム医療機器として製造販売承認を取得した。プログラム医療機器は遠隔医療にも適しており，国の DX 推進政策にも合致しているため厚生労働省や医薬品医療機器総合機構 PMDA も積極的に応援している。確かにプログラム医療機器は幅広い患者層に，簡便に，テーラーメイドな医療を提供できる可能性を秘める。一方で，医薬品開発に比較して承認審査に用いられるエビデンスは相対的に少ないため，製造販売後調査も含めて，その対象患者，効果，安全性に関するエビデンスの積み上げが望まれる。

精神疾患には睡眠障害が高率に併存し，認知・社会機能，QOL の低下や疾患の再発・再燃を高めることはよく知られている。Research domain criteria（RDoC）においても，睡眠覚醒と概日リズムは主要なドメインの構成概念として示されている。精神疾患患者の睡眠評価に際しては主観症状のみならず客観症状の把握が求められるようになっているが，近年続々と登場している種々のウェアラブルデバイスの多くは睡眠 / 覚醒の判定は比較的高い精度で判定できるものの，レム / ノンレム構造や睡眠深度などの睡眠構築となるとまだ判定合致率が十分ではなく改善の余地が大きい。そこで在宅で患者自身が装着可能な少数チャンネルの脳波計が開発され，以前よりも廉価となり，研究や臨床で用いられるようになってきた。脳波データは周波数パワーや紡錘波，レム睡眠特性など情報量（特徴量）が多く，生物学的精神医学研究でも今後大いに活用が期待されている。

ところが，ここで問題になるのが脳波データ解析のスループットである。トレーニングを積んだ臨床検査技師でも一晩分の睡眠ポリグラフの視察解析に 2 時間程度かかるため，日々多くの患者から取得される脳波データの解析がボトルネックとなっていた。しかし，最近では睡眠脳波データの AI 解析能力が急速に向上している。筆者らが企業と共同で取り組んでいるのは，睡眠脳波の周波数パワー値を画像データ化して畳み込みニューラルネットワークで機械学習させ，レム / ノンレム睡眠や睡眠深度を見分ける判読システムの構築である。

現時点でもそれなりの判定精度が得られているが，臨床で広く応用するにはもう少し性能を上げたい。特に精神疾患患者の睡眠脳波は疾患独自の，そして向精神薬の影響で健康な人のそれとは異なる特徴をもつので判定が難しい。そのため多数の患者の睡眠脳波を教師データとすべく収集したのだが，その際に思いがけない問題が露呈した。当初，ある経験値の高い臨床検査技師の判定結果を「正解」として機械学習させたのだが，念のために同じ睡眠脳波データを複数の熟練技師に判定してもらったところ，「人－人」間での判定合致率が想定以上に低く，「人－AI」間の判定合致率とさほど違わなかったのである。要するに「（教師データの）正解とは何か？」というグレーゾーン診断のラベルノイズ問題に直面している。ラベルノイズ問題はどの研究領域でもあるとはいえ，睡眠脳波（おそらく臨床脳波も）ではグレーゾーンが些か広すぎる。しかし少しうがった見方をすれば現在の睡眠脳波の AI 診断能力は人の診断の曖昧さレベルに近づいたともいえる。さて「正解」をどのように定義するか，また診断精度の目標をどこに設定するか，実に悩ましい。