機械学習を用いた診療イベント予測

キヤノンメディカルシステムズ㈱　研究開発センター　先行技術研究部
杉山　真哉

【はじめに】

1. データ利活用

医療現場のニーズは時代と共に常に変化している。当社ではそのニーズに応えるため、従来提供してきた医用画像診断機器群に加え、ヘルスケア IT に注力している。画像を含めた多様な医療情報を収集・統合・分析・加工し、より付加価値の高いソリューションを提供していく(図 1)なかで、それらを支えるテクノロジーの一つが機械学習をはじめとする AI 技術であると捉えている。

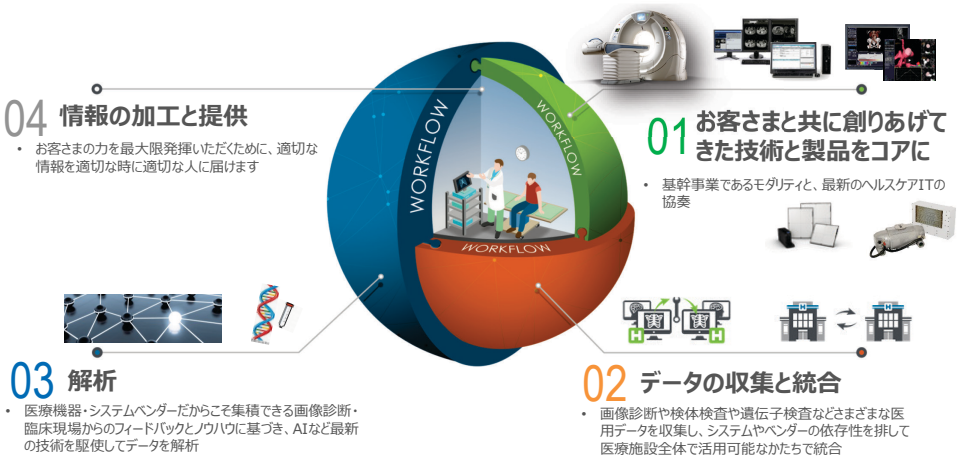

図1 医療情報の収集・統合・分析・加工

2. 診療意思決定支援とその課題

当社では医療におけるヘルスケア IT ソリューションを、「医療従事者が、より良い医療を効率よく患者に提供するための知恵」と位置付けている。

画像診断機器等から得られたデータは、そこから診療において活用される情報が抽出され、さらに診療ノウハウとして知識、さらに知恵へと変換・統合・意味付けされていき、各段階において、診療ワークフローにおけるさまざまな意思決定へと活用される(図 2)。その各段階における診療意思決定支援（Clinical

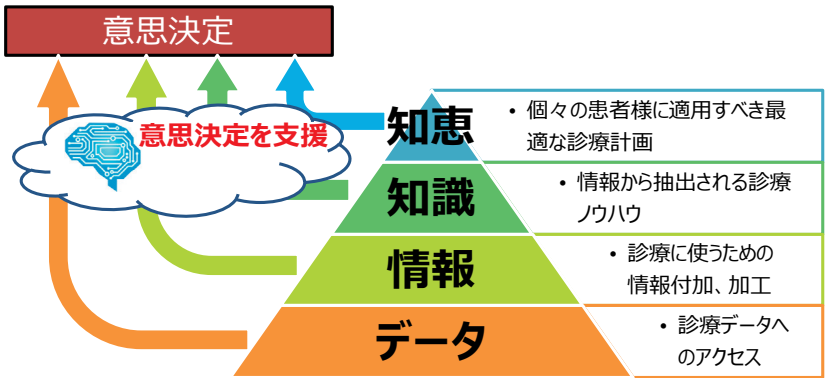

図2 診療意思決定支援におけるデータ・情報・知識・知恵の活用

Decision Support; CDS)が期待されており、ヘルスケア IT を用いて提供できるサービス・価値である。

CDSを実現するには多くの技術課題があるが、特に、データ種による取り扱い、機械学習によるモデル構築、モデルの有用性評価、という3点に注目している。当社が実施した「小児循環器領域における機械学習による診療イベントの予測」[1)]の研究を事例に、その技術課題と取り組みを解説する。

【小児循環器領域における機械学習による診療イベントの予測】

1. 研究概要

(1) 背景・目的

小児循環器領域では循環動態の管理が診療の大きな割合を占める。循環動態が悪化し急性心不全となると、緊急の手術や長期間の投薬が必要となる場合もあり、患者の予後に大きな影響を及ぼしうる。急性心不全の予測支援が実現すれば、初期対応の迅速化や、重症化の防止に貢献することが期待されている。

我々は、これまでも小児循環器領域で機械学習を用いた診療イベント予測モデルの精度を向上させる特徴量およびアルゴリズムの検討を行なってきた[2,3)]。本研究では、診療イベントを予測する機械学習モデルと、その予測結果および予測根拠を適切なタイミングで提示する診療イベント予測表示システムを構築する。構築した予測モデルの精度、予測表示システムの使用性(Usability)および説明可能性(Explainability)の評価を行ない、診療意思決定における支援の可能性を明らかにする。

(2) 方法

本研究における予測対象の診療イベントは、入院中の急性心不全の発生とした。ここで、急性心不全の発生の定義は、(1)心臓関連の緊急手術の実施、(2)侵襲度の高い治療的介入(体外式膜型人工肺、呼吸器管理)の実施、(3)心不全治療薬であるカテコラミンの投与および種類の追加、のいずれかに該当するものとした。予測のタイミングは、治療判断を変えることで急性心不全を防止できる可能性が高い2日以内とした。

2015 年 1月から 2017 年 12月までに小児循環器疾患で入院した患者のデータを収集し、7 歳未満の患者のデータのみを用いて循環動態悪化の予測モデル作成および精度検証を行った。

診療イベント予測モデル構築に用いる対象患者のデータは、予測モデルの学習用、検証用およびテスト用に患者を分割し、教師あり学習を行なった。機械学習への入力には、バイタルサイン等の時系列数値データなど 65 種類の特徴量を用い、出力は予測される診療イベントの有無とした。予測精度の検証には、学習に使用したデータとは別のテスト用データを用いた。

作成した予測モデルは、電子カルテと連携した診療イベント予測システムとして試行的に組み込んで有用性を評価した。予測表示には、「アラート画面」および「患者個別画面」の 2 画面を用意し、利用者には最初にアラート画面を提示した。アラート画面には複数患者の直近のアラート履歴を同一時間軸上に表示することで、どの患者でアラートが出ているか、アラートがどれくらい集中しているかを容易に把握できるようにした。また患者個別画面には、各アラートの根拠データや、アラート発生時付近の診療データを提示する詳細データエリアを設けた。ここで表示する根拠データは、各患者の予測結果に対する寄与率が高い特徴量に関係する診療データとした。

(3) 結果

475人のデータを用いて予測モデル作成および精度検証を行った結果、テストデータ(患者 93人、8,979サンプル)において AUC=0.88 で循環動態悪化を予測できた。感度 81.1%、特異度 79.7%となる予測確率でのテストデータでの混同行列を表 1 に示す。

診療イベント予測システムでは、循環動態悪化が予測される患者がリアルタイムに検出され、画面上にアラートとして通知した。また、胸部レントゲン像やバイタルサインなどの様々な診療データが一画面内で閲覧可能であるため、循環動態悪化予測の妥当性確認や次の治療の判断を行うことが可能となった。図3に示す診療イベント予測表示システムのアラート画面例では、画面上部には複数患者の直近のアラート履歴を同

一時間軸上に表示した。例えば、入院中の複数患者を対象に、アラートの有無と内容で対応の優先度を確認できる。画面下部の詳細データエリアでは、診療データのグラフを表示している。赤でハイライトされた領域で予測アラートの発生期間を示している。

表1 テストデータにおける混同行列
（最も長い期間入院した患者の入院日数に合わせて重み付けしたサンプル数）

		予　測	
		Negative	Positive
正　解	Negative	1,164,112	29,685
	Positive	3,789	16,307

図3 診療イベント予測表示システムのアラート画面

診療イベント予測システムでは、循環動態悪化が予測される患者がリアルタイムに検出され、画面上にアラートとして通知した。また、胸部レントゲン像やバイタルサインなどの様々な診療データが一画面内で閲覧可能であるため、循環動態悪化予測の妥当性確認や次の治療の判断を行うことが可能となった。図 3 に示す診療イベント予測表示システムのアラート画面例では、画面上部には複数患者の直近のアラート履歴を同一時間軸上に表示した。例えば、入院中の複数患者を対象に、アラートの有無と内容で対応の優先度を確認できる。画面下部の詳細データエリアでは、診療データのグラフを表示している。赤でハイライトされた領域で予測アラートの発生期間を示している。

2．診療イベント予測のための時系列データの取り扱い

医療におけるデータ種は数値、画像、自然言語など多種多様である。データから知識、さらに知恵へと変換・統合・意味付けしていくためには、それぞれのデータ種ごとに取り扱い方法を検討する必要があるが、本研究では特に数値の時系列データに注目した。

本研究では、急性心不全兆候のイベント予測に重要なバイタルサイン（呼吸数、心拍数、収縮期血圧、拡張期血圧、SpO_2）、検体検査（ヘモグロビン、NT-proBNP、血清クレアチニン、ラクテート、動脈血酸素分圧）、尿量、体重、月齢などの時系列データを、電子カルテから抽出した。

バイタルサインなどは、高頻度で取得される診療データであるが、測定は不定間隔であることが多く、取得状況によってその測定値の品質も大きく変化しうる。そのため入力である時系列数値データは、変動を前提として取り扱えるように前処理を行った。

例えば、時系列の一定区間（ウィンドウ）内の中央値や四分位範囲などの統計量を計算し、その統計量を予測における特徴量として用いた。統計量を用いることで、含まれるノイズの影響を軽減するだけでなく、データの種類毎に記録日時が異なる時系列データを扱うことができる。本研究ではウィンドウをスライディングさせて特徴量を算出し、ウィンドウの開始および終了時刻の取り方のバリエーションを設けた。特徴量算出のウィンドウ幅は24時間で、スライディング幅は8時間とし、合計65種類の特徴量を予測モデルに用いた。

また欠損値への対応として、最後に観測された値でその後の一定期間を補完するLOCF(Last Observation Carried Forward)補完を全ての特徴量に適用した。LOCF補完を適用する期間は最後に値が観測された時点から最大10日間とした。

予測モデル学習のためのサンプルの教師ラベルは、診療イベントが発生する2日前から当日にアラートを提示させる場合、3日前から1日前のウィンドウをPositive、それ以外のウィンドウをNegativeとした。

呼吸器管理・体外式膜型人工肺の実施中、および術後2週間は学習の対象外とした。

3．診療イベント予測モデル

本研究では「予測」を対象にモデル作成を行った。一般的に機械学習では、Descriptive(現状分析)、Predictive(将来予測)、Prescriptive(対策推奨)といった3つの分野で機能を提供できる。それらは、どのようなCDSを実現するか、すなわちどのような臨床意思決定を支援するか、という課題設定によって使い分けられる。

予測モデルのアルゴリズムはランダムフォレストを採用した。予測根拠説明のための寄与率算出にはLIME(Local Interpretable Model-agnostic Explanations)[4)]を利用した。特徴量の算出および予測モデルの学習・評価には統計ツールRを用いた。

予測モデルの学習、検証およびテストには、475人のうちPositiveあるいはNegativeラベルが1つ以上付与された460人を用いた。学習データ（患者267人、29068サンプル）でランダムフォレストモデルを学習させ、検証データ（患者100人、11563サンプル）でパラメータを調整した結果、木の深さ5、木の数100のランダムフォレストモデルが得られた。テストデータ（患者93人、8979サンプル）におけるランダムフォレストモデルを精度評価し、AUC=0.88の結果を得た。

患者状態の変化によるノイズの影響が懸念されたが、特徴量を統計値にしたことにより、小児特有のセンサ外れや啼泣などによるノイズの影響を軽減できたと考えられる。対象患者である小児患者は月齢によりバイタルサインの基準値が異なる[5)]ため、非線形モデルであるランダムフォレストを選択したことが高いAUCに繋がったと考えられる。

4．有用性評価

診療意思決定支援においては、予測モデルの精度だけでなく予測結果をどのように提示するかという点も重要となる。予測結果を診療における意思決定に結び付けるためには、予測の根拠や過程を提示したり、適切なタイミングで予測機能を提供したりすることになる。すなわち、CDSを評価するうえでは、予測モデルの精度だけではなく、予測表示システムの使用性や説明可能性もその対象とする必要がある。予測表示システムが高い説明可能性を持っていれば、偽陽性が多い場合でも利用者自身でその真偽を判断でき、次に行う診療の意思決定の検討を円滑に進めることができる可能性がある。

診療イベント予測表示システムの評価は、使用性および説明可能性に関するアンケート評価で行なった。それぞれ複数の質問を用意し5段階で評価した。循環器内科の医師3名は、評価期間として定めた7日間のうちいずれか2日間で予測表示システムを利用し、アンケートを記入した。アンケートの評価項目について、使用性はソフトウェア品質の国際規格 ISO 9126における使用性およびウェブサイトユーザビリティアンケート評価手法の先行研究[6)]を参考に作成し、評価項目は理解性、習得性および運用性の3つとした。説明可能性の評価項目はアメリカ国防高等研究計画局（Defense Advanced Research Projects Agency; DARPA）による説明可能なAIの効果測定指標[7)]を参考に作成し、評価項目は予測根拠説明の満足度、

メンタルモデル、タスクパフォーマンスおよび信頼評価の 4つとした。

医師 3名が予測表示システム利用後に回答したアンケートの結果を表2 に示す。使用性の全ての評価項目(理解性、習得性および運用性)において、5 段階評価の「やや良い」を示す平均 4.0を獲得した。説明可能性については、「どちらとも言えない」から「やや良い」を表す平均 3.0～3.6となった。

メンタルモデルの質問は 5つあり、そのうち「次に取るべき行動の利用者自身による理解の程度」に関する質問の回答は平均 4.7と高かった。タスクパフォーマンスに関する質問は 4 つで、全ての回答の平均は 3.1だったが、そのうち予測表示システムの利用による「診療判断の前倒し」および「診療判断の時間短縮」への効果を尋ねる質問の回答は平均 4.0となった。

診療イベント予測表示システムの使用性のうち、理解性および習得性の高さには、シンプルな画面構成が理解性および習得性を高めたと考えられる。評価に用いた予測モデルは感度を優先して 1日あたりのアラート患者数を 4.4人と多く設定したため、誤検出による診療判断への悪影響を懸念していたが、予測表示システムのアンケート評価では診療判断の前倒しおよび時間短縮に繋がる可能性が示唆された。アラート画面の表示方法により、利用者自身がその真偽を容易に判断でき、感度を高く設定しても「アラート疲れ」が起こりにくくなったためと考えられる。

一方で、説明可能性の満足度は「どちらとも言えない」を示す平均 3.0 という結果だった。今回採用したランダムフォレストは複雑なモデルであり、決定木などの他のモデルと比較して一般的に説明可能性が低い。利用者が想定していなかったパターンで予測モデルが予測できた場合、特徴量などの具体的な予測モデルの根拠説明があったとしても、利用者はその妥当性を判断できない場合がある。利用者が普段用いる言葉や基準に変換したり、従来の予測と異なるパターンであることを説明したりすることで、満足度を改善できる可能性があると考えられる。

表2 アンケートの回答結果 (5段階評価)

対　象	項　目	回答(平均)
使用性	理解性	4.0
	習得性	4.0
	運用性	4.0
説明可能性	満足度	3.0
	メンタルモデル	3.6
	信頼評価	3.3
	タスクパフォーマンス	3.1

【まとめ】

小児循環器領域における機械学習による診療イベントの予測を事例に、CDS実現の取り組みにおける注目点を解説した。取り組みのなかでは特に、データ種による取り扱い、予測モデル構築、有用性評価に注目した。

多様なデータ種を取り扱えるようにするには、データ種に合わせた取り組みが必要である。研究事例としては時系列データについて解説したが、画像データでは画像特徴量を抽出することで他のデータ種、例えば放射線量などと融合して解析する研究を進めており[8,9]、自然文として記載されたデータについても構造化することで機械学習に適用する研究を実施している[10]。医療情報の特徴である多種多様なデータ種を適切に取り扱うことは、解析に取り組む基盤として最も重要な要素である。

機械学習を用いた解析では、現状分析か、予測か、それとも対策推奨か、目的に応じて手法を使い分けることが必要となる。また、目的に応じて、用いるべきデータの時間的条件、空間的条件(施設など)、評価指標(出力)定義が異なる。それらは、診療においてどのような意思決定を支援するかという課題定義と直結し

ている。

本研究ではアンケートをベースに有用性評価を行ったが、さらに定量的な評価手法等を検討していく必要があり、説明可能性の向上を目指していきたい。精度検証だけでは確認できない場合が多いため、今後CDSの有用性評価手法について、統一された方法論の確立や評価指標の作成が議論され、学術的にもさらなる検討、発展があることを期待している。

また、CDSを臨床に適用することを想定すると、例えば診療医へ予測アラートを通知すると同時に、関連する画像やバイタルサインも俯瞰的に表示することが適切と考える(図4)。診療意思決定支援は診療フローにおける重要ステップに焦点を絞って支援し、統合表示は診療フロー全体の俯瞰を担うことで、効果的、効率的な診療支援を多面的に実現できる。

医用画像および医療情報を活用し、医療従事者と患者のために「より充実した医療に向かうための知恵」を集約し、それに基づいた数々のソリューションを迅速に提供し続けていくことによって「Made for Life」の具現化を目指している。

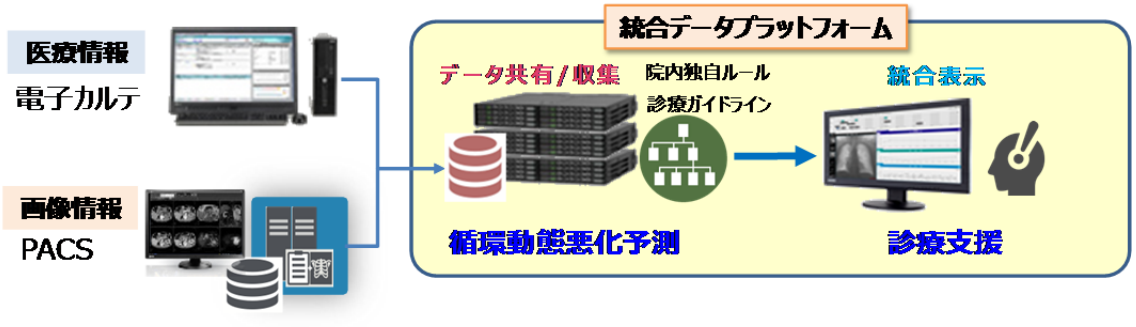

図4 診療支援と統合表示

【参考文献】

1) 佐藤 杏莉、狩野 佑介、下西 健太ら. 機械学習を用いたリアルタイム診療イベント予測表示システムによる診療支援の可能性 - 小児循環器領域の場合-. 医療情報学会連合大会 2019.

2) 若宮卓也, 野木森宜嗣, 加藤昭生ら. データの変化をとらえる診療支援システムを用いた、心不全の予測. 日本小児循環器学会、2018.

3) 若宮卓也, 杉山隆朗, 野木森宜嗣ら. 診療支援システムを用いた急性心不全の予測 ～第 2 報～. 日本小児循環器学会、2019.

4) Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. KDD 2016.

5) Kliegman RM, Stanton BF, St Geme III JW et al. ネルソン小児科学 原著第 19 版. エルゼビア・ジャパン, 2015.

6) 仲川薫, 須田亨, 善方日出夫ら. ウェブサイトユーザビリティアンケート評価手法の開発. 第 10 回ヒューマンインターフェース学会紀要, 2001.

7) Gunning D. Explainable Artificial Intelligence (XAI) DARPA/I2O Program Update November 2017.DARPA,2017.[https://www.darpa.mil/attachments/XAIProgramUpdate.pdf(cited 2019- Aug-20)].

8) M. Nakatsugawa, Z. Cheng, K. Goatman et al. "Radiomic analysis of salivary glands for the prediction of weight loss in irradiated head and neck cancer patients" Radiological Society of North America. 2016.

9) M. Nakatsugawa, Z. Cheng, K. Goatman et al. "Radiomic analysis of salivary glands and its role for predicting xerostomia in irradiated head and neck cancer patients", International Journal of Radiation Oncology Biology Physics Vol. 96, 2016.

10) P. Rogalla, B. E. Hoppel, K. Noro et al. "Automated CT Abdominal Imaging Protocol Selection Using a Natural Language Processing and Machine Learning" European Congress of Radiology. 2019.