

一般口演 | 医療支援

一般口演17 医療支援

2019年11月24日(日) 09:00 ～ 10:00 A会場 (国際会議場 2階コンベンションホールA)

[4-A-1-03] 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～

○佐藤 杏莉¹、狩野 佑介¹、下西 健太²、上田 秀明³、杉山 真哉¹（1. キヤノンメディカルシステムズ株式会社 研究開発センター, 2. キヤノンメディカルシステムズ株式会社 ヘルスケアIT第二事業部, 3. 神奈川県立こども医療センター 循環器内科）

キーワード：Clinical Decision Support, Prediction, Usability, Explainability, Machine Learning

【背景】近年、疾患の発症や重症化、急変といった診療イベントのリスクを評価し、初期対応やコントロールに貢献することが期待されている。診療イベントの予測ではしばしば性能や精度が意識されてきたが、説明可能性にも注目が集まり始めている。

【目的】診療イベントを予測するモデルを開発し、その予測結果と予測の根拠をリアルタイムで提示するシステムを構築、有用性を検証することである。

【対象】予測モデル開発の対象患者は、2015-2017年の3年間に先天性心疾患で入院した7歳未満の患者475人である。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。

【方法】電子カルテから抽出したバイタルサイン、検体検査、尿量および体重の過去データから時系列変化に関する特徴量65種類を計算し、2日以内に急性心不全の発生（以下、アラート）有無を予測するランダムフォレストモデルを開発した。また、電子カルテおよび PACS と接続し、予測モデルを用いてアラート有無を判定する診療支援システムを開発した。システムには、入院患者のアラート履歴をタイムライン上に一覧表示し、予測モデル計算に用いたデータをアラートの根拠として表示した。患者個別画面では種々の診療データを統合的に閲覧できるようにした。

【結果】過去データについてアラート有無を AUC=0.87 で予測できた。診療支援を想定したアラート表示を行うシステムを構築することができた。

【考察】機械学習を用いた診療イベントの予測では、誤検出を完全になくすることは困難なため、簡便に予測の妥当性を判断できる提示上の工夫が診療支援システムでは重要である。

【結論】機械学習モデルにより予測した診療イベントのアラートおよび根拠データの、リアルタイム表示による診療支援の可能性が示された。

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 -

佐藤 杏莉^{*1}、狩野 佑介^{*1}、下西 健太^{*2}、
上田 秀明^{*3}、杉山 真哉^{*1}

*1 キヤノンメディカルシステムズ株式会社 研究開発センター、

*2 キヤノンメディカルシステムズ株式会社 ヘルスケア IT 第二事業部、

*3 神奈川県立こども医療センター 循環器内科

Clinical Event Prediction with Machine Learning for Clinical Decision Support - A Case of Pediatric Cardiology -

SATO Anri^{*1}, KANO Yusuke^{*1}, SHIMONISHI Kenta^{*1}, UEDA Hideaki^{*2}, SUGIYAMA Shinya^{*1}

*1 Canon Medical Systems Corporation, *2 Department of Pediatric Cardiology, Kanagawa Children's Medical Center

Predicting adverse clinical events leads to preventing patients from severe condition. In the prediction of clinical events by machine learning, not only accuracy but also usability and explainability in a clinical decision support (CDS) system are crucial because they help to understand and trust the model, and to improve the effectiveness. This research aims to develop a machine learning model to predict clinical events, to implement a system that provides the prediction results and the explanations, and to evaluate the usefulness. An acute heart failure predictive model was developed with records of 475 patients who were hospitalized for congenital heart diseases from 2015 to 2017. We calculated 65 features from numeric time-series data extracted from electronic medical records (EMR) and constructed a random forest model. The adverse clinical events were predicted at adjusted AUC = 0.88. We developed a CDS system that connects to EMR and PACS, and used the predictive model to detect the sign of adverse clinical events. In the system, the alerts were listed on the timeline, and the data contributed to the prediction were displayed as the explanation. By evaluating the accuracy, usability and explainability, the possibility of CDS with a machine learning model was shown.

Keywords: Clinical Decision Support, Prediction, Usability, Explainability, Machine Learning

1. 背景

病院内に蓄積された大量の診療データを利用することで、臨床的有用性の高い診療意思決定支援 (Clinical Decision Support; CDS) を実現することが期待できる。CDS の適用先の 1 つとして、疾患に起因する患者状態の急変や治療による副作用の発生といった診療イベントの事前予測が考えられ、初期対応の迅速化や、見逃しによる重症化の防止に貢献することが期待されている。

機械学習を用いた CDS は、医用画像データを用いた診断支援分野で最も進んでいる。医用画像はデータ取得頻度や回数は多くないものの、一定の水準の品質が担保されているという特長があり、機械学習で活発に取り扱われている。

一方で、診療イベントを予測する CDS では、バイタルサインなど、より高頻度で取得される診療データが重要となる。これらのデータの測定は不定間隔であることが多く、またデータが取得された時の患者の状況によってその測定値の品質も大きく変化する。そのため、変動を前提とした状況において適用可能な予測モデルの構築方法が求められる。

また、機械学習を用いた CDS を臨床で用いることを想定すると、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 (Usability) や説明可能性 (Explainability) 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、次に行な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。

我々は、これまで小児循環器領域で、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ってきた^{1) 2)}。本研究では、診療イベントの予測モデルの精度に加え、診療イベント予測表示の使用性と説明可能性の評価を行ない、システムとしての診療意思決定支援の可能性について検討する。

2. 目的

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および説明可能性の評価を行ない、予測表示システムを実際の臨床現場で使用することを想定した場合の診療意思決定支援の可能性を明らかにする。

3. 方法

本研究では、疾患や診療シーンを限定しない場合の一般的な診療イベントに対して診療イベント予測モデルおよび診療イベント予測表示システムを構築する方法を検討し、それを小児循環器領域に適用することで予測モデルおよび予測表示システムを構築、評価した。本評価は神奈川県立こども医療センター倫理委員会に承認された方法で行なわれた。

3.1 小児循環器領域における診療イベントの抽出

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。しかしながら、対象が小児であるがゆえにその体調の変化は患者自身の言葉で説明されないことがほとんどで、収集される診療データや患者の様子などを継続的に観察することから循環動態を知るしかない。さらに、後から病状やカルテを振り返ればその前兆を認めていることもあるが、実際にはその場では気付けないことも多い。急性心不全の予測支援が実現すれば、重症化の防止に大きく貢献できると予想される。

そこで、本評価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。

3.2 診療イベント予測モデルの構築

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。

診療イベントの予測をする機械学習モデルの入力にはバイタルサイン等の時系列数値データを用い、出力は予測される診療イベントの有無とした。

入力である時系列数値データは一定区間(ウィンドウ)内の中央値や四分位範囲などの統計値に処理され、その統計値を特徴量として用いた。統計値を用いることで、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。また、記録頻度が低いデータへの対応、すなわち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。さらに、特徴量はウィンドウをスライディングさせて算出し、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増やした。ウィンドウの開始から終了までの時間幅をウィンドウ幅、スライディングの時間幅をスライディング幅と呼ぶ。

サンプルは統計値処理をしたウィンドウ1つに対して1つとし、それぞれのサンプルにラベルを付与した。学習データのラベルは、診療イベントが発生するX日前から当日にアラートを提示させる場合、(X+1)日前から1日前のウィンドウをPositive、それ以外をNegativeとした。ただし、ウィンドウをスライディングする場合はPositiveおよびNegativeが重なるウィンドウが現れるため、そのウィンドウは学習の対象外とした。

個々のサンプルに対する予測モデルの予測根拠の提示(局所的な説明)を行なうために、予測結果ごとに各特徴量の寄与率を算出した。

3.2.1 小児循環器領域における診療イベント予測モデルの構築と評価

診療イベント予測モデルの構築の対象患者は、2015年1月1日から2017年12月31日までに神奈川県立こども医療センターの循環器内科あるいは心臓血管外科に先天性心疾患で入院した7歳未満の患者のうち、目的が検査入院でない患者とした。

電子カルテからバイタルサイン(呼吸数、心拍数、収縮期血圧、拡張期血圧、SpO₂)、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、月齢および入院開始からの経過時間の時系列データを抽出した。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、LOCF補完を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。合計65種類の特徴量を予測モデルに用いた。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。

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、予測根拠説明のための寄与率算出にはLIME(Local Interpretable Model-agnostic Explanations)³⁾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。

診療イベント予測モデルの評価には、各患者のサンプル数が等しくなるよう入院日数に応じてサンプルの重みを調整して導出したAUROCの値(以下、補正AUC)、感度および特異度、1日あたりのアラート患者数を用いた。

3.3 診療イベント予測表示システムの構築

予測モデルへの入力データをリアルタイムで収集するため、診療イベント予測表示システムは電子カルテなどの院内システムと接続した。収集したデータは予測表示システムに蓄積し、予測モデルの計算および画面でのデータ表示に用いた。

診療イベント発生が予測される場合の利用者への提示方法には、アラート音やメールなどで利用者が受動的に情報を受け取るプッシュ型が一般的であるが、感度を高くなるように設定した場合には偽陽性が多くなり、利用者のアラートへの反応が鈍くなる「アラート疲れ」が問題となる。そこで我々は、利用者が能動的に予測表示システムにアクセスした際に結果を提示するプル型を採用するとともに、偽陽性が多い場合でも利用者自身でその真偽を容易に判断できる提示方法を用いた。

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、アラート画面にはアラート履歴だけでなく、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、アラート画面上で利用者自身が偽陽性アラートを排除できるようにした。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。

利用者がアラート画面に表示される複数患者の中から一人の患者を選択した際には、患者個別画面へと遷移し、根拠データだけでなく種々の診療データを統合的に提示した。これにより、予測の真偽判断だけでなく、次にどのような診療判断をすべきかを容易に検討できるようにした。

3.3.1 小児循環器領域における診療イベント予測表示システムの構築と評価

診療イベント予測表示システムは2018年12月1日から2019年6月10日まで電子カルテおよびPACSと接続し、循環器内科あるいは心臓血管外科に入院中の患者について1日1回データを転送し、そのうち7歳未満の患者に対して急性心不全の予測を行なった。予測表示システムでは、特徴量

計算に用いられたデータ種類およびアラートに加えて、カルテ記載、看護記録、画像検査、投薬および患者基本情報を表示できるようにした。なお、予測表示システムはサーバ(Windows Server 2012 R2)およびクライアント(Windows 10)で構成され、サーバ上でデータ収集および予測モデルによる診療イベント予測を行ない、クライアント上のウェブブラウザでアラート画面および患者個別画面を表示した。

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。5段階評価は1が最も悪く、5が最も良くなるように設定し、各評価項目で平均値を算出した。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⁴⁾を参考に作成した。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⁵⁾を参考に作成した。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。予測根拠説明の満足度は、予測表示システムによる急性心不全の判断根拠の説明について、その明瞭性および利用者へ与える自信の程度を測った。メンタルモデルで測定するのは、予測モデル全体や個々の予測結果に対する利用者の理解、次取るべき行動の利用者自身による理解の程度とした。タスクパフォーマンスでは、予測表示システムを利用することにより意思決定の正確性や効率性が向上する程度、利用者の理解度に応じて予測表示システムの根拠説明の方法を変えていると感じる程度を測った。信頼評価での測定は、予測表示システムを信頼できるケースを利用者が理解している程度とした。

4. 結果

4.1 急性心不全予測モデル構築の対象データ

急性心不全予測モデルの構築に用いた対象患者は475人であった。急性心不全が1度以上発生した268人を急性心不全患者、全く発生しなかった207人を非急性心不全患者とし、患者属性を比較した(表1)。月齢は、対象期間内の初回入院開始時点での月齢で、平均±標準偏差を示した。全ての属性においてウェルチのt検定あるいは χ^2 検定による両群の有意差は認められなかった。

表1 対象患者の属性

属性	急性心不全 (n = 268)	非急性心不全 (n = 207)	p
男児 [%]	53.7	54.1	0.935
月齢 [ヶ月]	10.5 ± 17.7	11.7 ± 19.8	0.501
染色体異常 [%]	14.9	15.5	0.872
内臓心房錯位 [%]	6.7	3.9	0.172

4.2 急性心不全予測モデルの評価結果

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の精度評価では、急性心不全の有無を補正 AUC=0.88(補正前 AUC=0.78)で予測できた(図1)。感度81.1%、特異度79.7%となる予測確率を4.3節の予測表示システム評価時の急性心不全の発生有無を判別する予測確率の閾値として採用した。このときのテストデータでの混同行列を表2に示す。ここで、混同行列の数値は、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を意味する。

予測確率の閾値を変化させたときの1日あたりのアラート患者数を横軸に、感度および特異度を縦軸にとると、先に設定した感度81.1%のとき、1日あたりのアラート発生患者数は4.4人となった(図2)。ただし、1日あたりの予測表示システムの評価対象患者数は、過去データから算出された平均患者数より16人と仮定した。

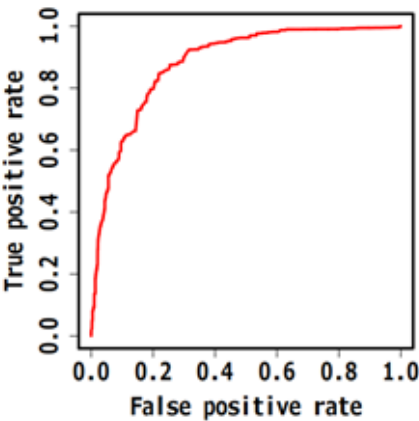


図1 ROC 曲線

表2 テストデータにおける混同行列

		予測	
		Negative	Positive
正解	Negative	1164112	29685
	Positive	3789	16307

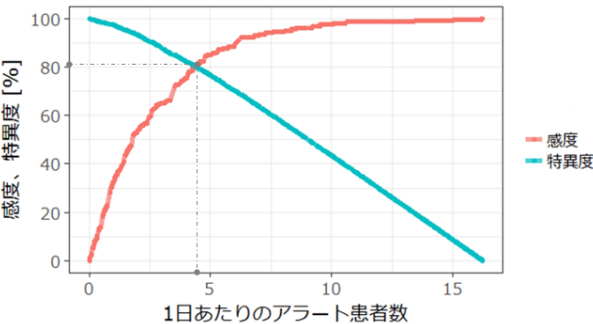


図2 アラート患者数と感度および特異度の関係

4.3 診療イベント予測表示システムの評価結果

予測表示システムを電子カルテおよびPACSと接続して急性心不全の予測を行なった結果、予測が適用された患者は119人、9224サンプルとなり、補正 AUC=0.76(補正前 AUC=0.70)となった。

評価期間である7日間で、対象診療科における7歳未満の入院患者は45人、そのうちアラートが発生した患者が35人となった。利用者がアクセスしたときに結果を提示するプル

型で表示したアラート画面および個別患者画面を、図 3 および図 4 に示す。図 3 のアラート画面上では、複数患者のアラートが確認でき、その根拠データや診療データを閲覧できた。図 4 の個別患者画面では患者の種々の診療データやアラートを統合的に提示できた。

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 3 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「次を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。

表 3 アンケートの回答結果

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

5. 考察

急性心不全を 2 日以内に予測するランダムフォレストモデルを構築し、テストデータでの補正 AUC が 0.88 となった。入力バイタルサインデータ等であり、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によるノイズも含めて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⁶⁾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。

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、画面構成が関係していると考えられる。予測表示システムでは画面が 2 つ(アラート画面、患者個別画面)で、それぞれの画面のエリアは両方で類似しており、上部にタイムラインを、下部に詳細のデータを表示するエリアとしていた。この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。また、プル型のアラートを採用したことで従来電子カルテを利用していたタイミングで予測表示システムを利用し、従来の診療ワークフローを大きく変更せずに取り入れることができたため、運用性の評価が高くなったと考えられる。しかしながら、プル型の場合、診療イベント発生に対してアラートの発見が間一髪になる、あるいは手遅れになる恐れがあり、対象とする診療イベントがプル型に相応しい緊急度であるかどうかをよく検討する必要がある。

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。開発構築し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」が起りにくくなったためだと考えられる。

使用性は高く評価されたものの、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。また、根拠データの提示が診療データであったため、統計値に基づいて発生したアラートの理由として説明可能性が低かったと考えられる。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。この問題に対して予測表示システムでは、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。

6. 結論

バイタルサイン等の時系列数値データを使った診療イベントの予測として、本研究では、小児循環器領域において急性心不全を予測するモデルを構築し、高い精度を達成することができた。構築した予測モデルを診療イベント予測表示システムに組み込み、その使用性および説明可能性をアンケートで評価した。本評価を通し、急性心不全の CDS は診療における意思決定の前倒しおよび時間短縮に効果があること、また高い運用性を期待できることを確認した。説明可能性は今後の課題であり、向上させることで利用者が CDS を信頼し、高い満足度で利用できることが期待できる。

参考文献

1) 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.

2) 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.

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

4) 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.

5) 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)].

6) 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.



図3 診療イベント予測表示システムのアラート画面

画面上部の時間軸上には、アラートが発生した患者名がアイコンに表示される。同一患者でアラートが複数回発生している場合にはアイコンは1つに集約され、その回数がアラート横に数字で表現される。画面下部にはアラートの根拠データを表示する詳細データエリアがあり、1つのボックスは1患者に対応する。このエリアには、各アラートに対して寄与率上位の診療データのグラフを表示し、グラフ上で濃い赤でハイライトされた領域はアラート発生期間を意味する。例えば、画面右下の患者の最新のアラートでは、尿量、拡張期血圧、呼吸数、体重および心拍数が寄与率上位となっており、それらの診療データのグラフが箱ひげ図として表示されている。患者を選択した場合にはその患者の患者個別画面に遷移する。



図4 診療イベント予測表示システムの患者個別画面

画面上部はその患者の診療イベントの経過を、画面下部には診療データを種類別に表示する。各パネルで赤くハイライトされた領域はアラート発生期間を意味し、アラート発生時付近の診療データを統合的に確認することができる。