一般口演 | 医療データ解析

一般口演1
医療データ解析

2019年11月22日(金) 09:00 ～ 11:00 C会場 (国際会議場 2階国際会議室)

[2-C-1-03] 機械学習と R/shinyを用いた患者個別の予測生存曲線描出アプリケーションの開発

○岡村 浩史[1]、中前 美佳[1]、橋本 匡生、森口 慎[1]、谷澤 直[1]、木村 友美[1]、中井 久実代[1]、田垣内 優美[1]、林 哲哉[1]、酒徳 一希[1]、井戸 健太郎[1]、原田 尚憲[1]、康 史朗[1]、南野 智[1]、中嶋 康博[1]、康 秀男[1]、中根 孝彦[1]、廣瀬 朝生[1]、中前 博久[1]、島崎 伸裕[2]、平松 武士[2]、下岸 亮祥[2]、 下野 直美[2]、竹本 恭彦[2,3]、日野 雅之[1,2] （1. 大阪市立大学 血液腫瘍制御学, 2. 大阪市立大学　医療情報部, 3. 大阪市立大学　総合医学教育学）

キーワード：shiny, machine learning, personalized survival prediction curve, random survival forest

【背景】臨床現場において各患者に対する予後予測のインフォームドコンセントは、既報告結果を踏まえた上でなされている。しかし、既報告が示す予後はリスク別・治療別に層別化されたものであり、背景や病歴が様々なリアルワールドの患者個別の要因に対して調整された予後予測を提案するツールはない。我々は過去の患者情報を csvファイルで入力することで、自動的に機械学習による予後予測モデルが作成され、新規患者に個別化された予測生存曲線を interactiveに描出することができる Webアプリケーションを開発した。

【方法】機械学習モデルは Random Survival Forestを用いた。 Webアプリケーション開発には R/Shinyパッケージを使用した。ユーザーが指定した outcomeと説明変数を用いて、治療介入日から10年後までの期間で自由に患者個別の予測生存曲線がプロット可能な仕様とした。試用データとして、当院の同種造血幹細胞移植の患者データを用いた。

【結果】過去の同種造血幹細胞移植患者のデータから予測モデルを作成し、さらにユーザーが新規患者の背景・移植法を入力することで、患者個別の要因によって調整された予測生存曲線を容易に描くことが可能であった。また、その予後予測精度は Harrell's c-index 0.72であった。*https://machine-learning-for-medicine.shinyapps.io/predictedEFS/*

【結論】この Webアプリケーションを利用することで、診療領域に限定されることなく、臨床医は過去の患者データを用いて新規患者個別の背景や治療介入法を反映した客観的な予後予測情報を容易に得ることができる。また、その結果をリアルタイムに患者に提示できるだけでなく、治療法別のシミュレーション結果を比較した上で、最適な治療法の選択を行うことができる可能性がある。

機械学習とR/shinyを用いた患者個別の予測生存曲線描出アプリケーション開発

岡村 浩史[*1]、中前 美佳[*1]、橋本 匡生、森口 慎[*1]、谷澤 直[*1]、木村 友美[*1]、中井 久実代[*1]、田垣内 優美[*1]、林 哲哉[*1]、酒徳 一希[*1]、井戸 健太郎[*1]、原田 尚憲[*1]、康 史朗[*1]、南野 智[*1]、中嶋 康博[*1]、康 秀男[*1]、中根 孝彦[*1]、廣瀬 朝生[*1]、中前 博久[*1]、島崎 伸裕[*2]、平松 武士[*2]、下岸 亮祥[*2]、下野 直美[*2]、竹本 恭彦[*2,3]、日野 雅之[*1,2]
*1 大阪市立大学 血液腫瘍制御学、*2 大阪市立大学 医療情報部、
*3 大阪市立大学 総合医学教育学

Web application for plotting personalized survival prediction curves using machine learning and R/Shiny

Hiroshi Okamura1[*1] Mika Nakamae[*1], Tadao Hashimoto, Makoto Moriguchi[*1], Nao Tanizawa[*1], Yumi Kimura[*1], Kumiyo Nakai[*1], Yumi Tagaito[*1], Tetsuya Hayashi[*1], Kazuki Sakatoku[*1], Kentaro Ido[*1], Naonori Harada[*1], Shiro Koh[*1], Satoru Nanno[*1], Yasuhiro Nakashima[*1], Hideo Koh[*1], Takahiko Nakane[*1], Asao Hirose[*1], Hirohisa Nakamae[*1], Nobuhiro Simasaki[*2], Takeshi Hiramatsu[*2], Akiyoshi Shimogishi[*2], Naomi Shimono[*2], Yasuhiko Takemoto[*2,3], Masayuki Hino[*1,2]
*1 Hematology, Osaka City University
*2 Medical Information System, Osaka City University
*3 Medical Education and General Practice, Osaka City University

Before clinicians obtain informed consent (IC) from a patient, they often present a patient-specific prognosis based on the previous similar patients' evidence. However, most prognosis information derived from previous evidence is stratified by a single predictor and no method exists in clinical practice to present a personalized survival prediction curve based on multiple prognostic factors. We have developed an interactive web application that plots a personalized survival prediction curve by using a machine learning model, regardless of the clinical field, when a user inputs a CSV file containing the training data set for machine learning into a web browser. We used random survival forest as the prediction model. The R/Shiny package was used to develop the interactive web application. We utilized the patients' data who underwent allogeneic stem cell transplantation (allo-HCT) at our hospital as trial data. Consequently, we could plot the personalized survival prediction curve for a new candidate of allo-HCT interactively using our web application. Thus, by inputting the training data set into this web application, clinicians in any medical field can obtain a personalized prognosis prediction adjusted for multiple predictors of a new patient. Furthermore, they can obtain IC by presenting an objective personalized prognosis prediction for their patients.

Keywords: shiny, machine learning, personalized survival prediction, random survival forest

1. 背景

臨床現場において各患者に対する予後予測のインフォームドコンセントは、疾患、疾患状態、治療法、予後スコア別などによって層別化された既報告結果を参考にした上でなされている。しかし、それら既報告が示す生存曲線はいずれか１つの予後因子によって層別化されたものであり、リアルワールドにおける多様な背景や病歴を有する患者個別の複数因子に対して調整された予測生存曲線を提案するツールはなかった。

近年、機械学習による予後予測モデルを作成し、個別の患者が有する複数の因子によって調整された予後予測を行うツールの報告がなされ始めている[1)]。しかしそれらは、特定の患者データによって予後予測モデルが既定されたものであり、研究者や臨床医、各施設や学会が保有する過去の患者データから、診療領域を問わず自由かつ直感的に予後予測モデルを作成し、患者個別の予後予測を行うことができるツールではない。

我々は、臨床医や施設・学会が各自で保有する過去の患者情報をCSV ファイルで入力することで、自動的に機械学習による予後予測モデルが作成され、新規患者に個別化された予測生存曲線を直感的かつ interactive に描出することができるGraphical User Interface（GUI）に基づいたWebアプリケーションを開発した。

2. 目的

診療領域を問わず、ユーザー（臨床医や各施設、学会など）が有する過去の患者データのうち、予後因子及び転帰情報から成るCSVファイルを入力することで、自動的にその患者情報を教師データとした予後予測モデルが作成され、さらに新規患者の予後因子情報を入力すると、その患者個別の生存曲線が描出される、GUIによるWebアプリケーションを開発する。

3. 方法

予後予測モデルはRandom Survival Forest（RSF）を用いた。RSFのアルゴリズムの要約は以下の通りである[2)]。

① 教師データからB個のブートストラップサンプルを作成する。平均 37%の症例はそれぞれのブートストラップサンプルに選ばれない。これらの症例をOut of Bag（OOB）と呼ぶ。
② それぞれのサンプル対するB個の生存木を作成する。各ノードではランダムに選ばれたmtry個の変数候補から、生存差が最大になる変数が選択される。
③ 各生存木に対する累積ハザード関数を算出し、それらの平均からアンサンブル累積ハザード関数を求める。
④ OOB データを用いて、アンサンブル累積ハザード関数の予測誤差を算出する。

変数重要度の計算には RSF の VIMP を利用した[3]。Web アプリケーション開発には R/Shiny パッケージを使用した。ユーザーが教師データである CSV ファイルにおいて指定した転帰と予後因子を用いて、治療介入日から 10 年後までの期間で自由に患者個別の予測生存曲線がプロット可能な仕様とした。

試用データとして、2008 年 1 月から 2017 年 11 月の間に当院で同種造血幹細胞移植を施行した 363 例の患者データを用い、アプリケーションの操作性、予測生存曲線の結果、及び予後予測精度を検証した。予測する転帰は全生存率とし、同種造血幹細胞移植の予後因子として報告されている年齢、performance status, 前処置強度、HLA 一致度、移植ドナーソース、refined Disease Risk Index (DRI-R), Hematopoietic Cell Transplantation Comorbidity Index (HCT-CI) の 8 つを特徴量とした。

RSF モデルの予測精度指標には 5 分割交差検証によって求められた各症例の予測死亡リスク値（predicted value）を用いて Harrell の c-index を算出した。c-index の 95%信頼区間算出には bootstrap 法（1,000 iterations）を用いた。

4. 結果

'https://machine-learning-for-medicine.shinyapps.io/predictedEFS/' において、当 Web アプリケーションが利用可能である。使用法は以下の通りである。

① 教師データである CSV ファイルを用意し、input 画面に Drag&Drop する（図 1）。

まず、次の仕様から成る CSV ファイルを用意する。CSV ファイルの行は症例を示し、列は予後因子または転帰の情報を示す。転帰情報の 2 列は列名が「Event」と「FollowUpDays」に固定されている。「Event」列は転帰を示す 0 と 1 から成る。0 は打ち切りを示し、1 は Event が生じたことを示す。「FollowUpDays」は打ち切りまたは Event 発症までの観察期間を示す。この 2 列以外の列全てはそれぞれ予後因子として扱われる。その CSV ファイルを input 画面に Drag&Drop すると、setting 画面に遷移する。

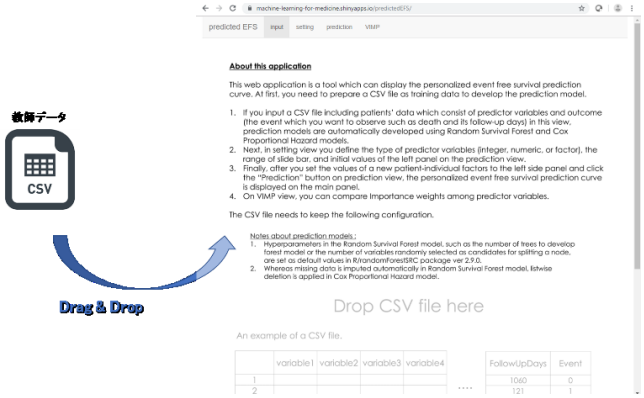

図 1 CSV ファイル入力画面（input 画面）

② setting 画面において prediction 画面の初期設定を行う（図 2）。

setting 画面では、各予後因子の型（integer, numeric, or factor）を定義する。予後因子が連続した整数または実数の場合それぞれ integer・numeric 型、カテゴリ変数の場合 factor 型を選択する。次の prediction 画面において、integer・numeric 型はスライドバーで、factor 型は select box で各因子を選択できる仕様となっている。

また、prediction 画面で調節する各因子の初期表示値を init 値、スライドバーの範囲を min、max 値として定義する。これらを定義した後、「prediction」タブを押すと、prediction 画面に遷移する。

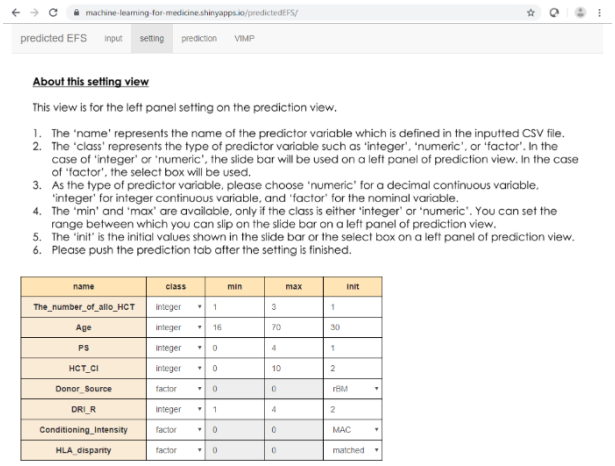

図 2 setting 画面

③ prediction 画面において、新規患者の情報を入力し、予測生存曲線をプロットする。（図 3）

prediction 画面の左パネルに、①②で規定した各予後因子がスライドバーまたは select box の形式で用意されている。ユーザーは、新規患者個別の因子をこれらスライドバーまたは select box に設定した後、「prediction」ボタンをクリックすると、その患者個別の予測生存曲線が main panel にプロットされる。またその際、RSF モデルの予測精度として OOB による Harrell の c-index も表示される。

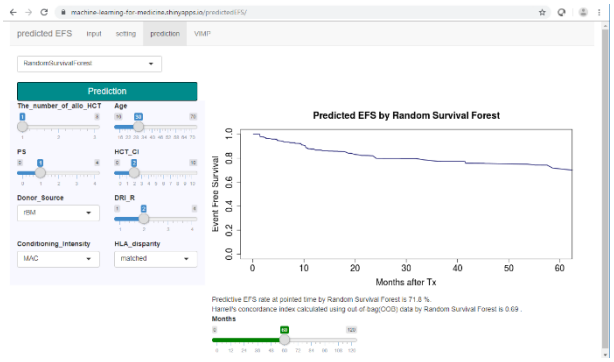

図 3 prediction 画面

④　VIMP 画面に転帰に対する変数重要度が描画される(図 4)。

転帰に対するそれぞれ予後因子の相対重要度を示すグラフが描画される。VIMP 値が大きいほど、RSF モデルにおいて転帰に対する重要度が大きいと評価された因子である。

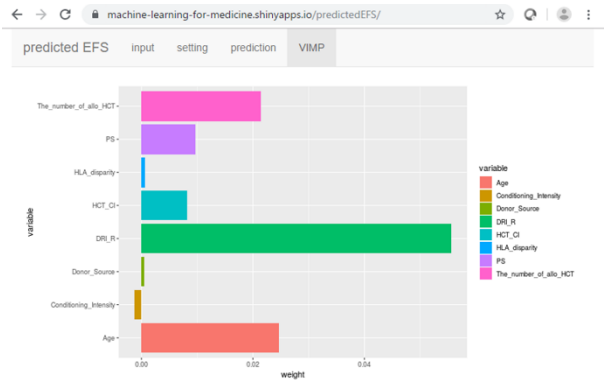

図 4　VIMP 画面

本アプリケーションの試用データとして、当院における過去の同種造血幹細胞移植患者のデータを入力 CSV ファイル（教師データ）として用いた。その結果、prediction 画面においてユーザーが新規患者の背景・移植法を入力すると、患者個別の複数要因によって調整された予測生存曲線を描くことが可能であった（図 3）。また、その予後予測精度は c-index：0.69 (95%信頼区間：0.65-0.72)であった。同一データによる Cox 比例ハザードモデルによる予後予測精度は c-index：0.72 (95%信頼区間：0.69-0.76)であった。

5. 考察

近年、医学研究領域において、機械学習を活用した研究論文が著増している（図 5：PubMed における”machine learning”の検索結果）。その中でも、診断または予後予測に関する研究が増加しており、様々な診療領域において機械学習による診断精度や予後予測精度についての検証がなされている。我々は今回、診療領域に依らずユーザー自身が保有する過去の患者データを入力することで、自動的に RSF による予後予測モデルが構築され、ユーザーが新規患者の情報を入力すると、interactive に患者個別の予測生存曲線を容易かつ直感的に描画することができる Web アプリケーションツールを開発した。

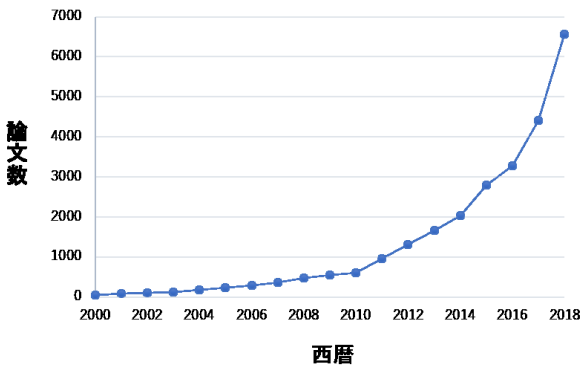

図 5　PubMed における”machine learning”の検索結果

試用データとして、当院で施行された同種造血幹細胞移植症例の患者データを教師データとして用いた結果、新規患者の予後因子を設定すると、患者個別の要因によって調整された予測生存曲線を描くことが可能であった（図 3）。これまで予後予測モデルを用い、患者個別の予後予測が可能なツールに関して、いくつかの報告がなされている [1]。しかしそれらのツールはいずれにおいても、既に特定の教師データから予後予測モデルが作成されたものであり、ユーザーが自ら保有している患者データから自由に予後予測モデルが作成できるものではなかった。我々が開発した Web アプリケーションは、ユーザーが教師データや予後因子、転帰を自由に選択することができる。そのため、教師データがあれば診療領域に依らず機械学習を用いた患者個別の予後予測が可能となる。また、これまでの予後予測ツールは、治療開始後 day180 の予後といったように、指定された点における予後予測を示すものであったが [1]、本アプリケーションでは患者固有の要因に合わせた時系列予測生存曲線を描くことが可能である。さらに、臨床医は本アプリケーションの prediction 画面に表示される予後予測モデルの予測精度指標 c-index を参考にし、それらの予後予測情報を実際の患者に情報提供することが妥当かどうか、を判断することができる。現在臨床現場において、患者個別の予後予測は既報告結果から得られる予後情報を主治医が患者固有の要因に対して経験的・主観的に調整し、伝えられることが多い。本アプリケーションを臨床現場に活用することで、医師はより客観的な予後予測情報を患者に伝えることができ、その上で治療選択の意思決定を行うことができるようになる。

機械学習モデルには、生存予測解析におけるアンサンブル学習モデルである RSF を用いた。RSF は教師データから複数の予後因子を統合した予後予測モデルを作り、新規患者の個別因子によって調整された生存曲線を描出することができる [4]。既存の一般的な生存解析モデルとしては、Cox 比例ハザードモデルがある。Cox 比例ハザードモデルを用いて、複数の予後因子を統合した予後予測モデルを作ることも可能であるが、Cox 比例ハザードモデルには比例ハザード性、または説明変数と転帰の間における対数線形性、といった前提条件が課せられる。さらに、交互作用を扱う場合には交互作用項を明示的に設ける必要がある。一方、RSF モデルは交互作用、非線形関係を自動的に扱うことができると共に、比例ハザード性のような前提条件を要しないノンパラメトリックなモデルであり、診療領域を問わず汎用的に患者個別の予測生存曲線を描画するツールの予測モデルとして望ましいと考えられた [5,6]。今回試用データとして用いた同種造血幹細胞移植症例の予後予測精度は、RSF に比べて Cox 比例ハザードモデルの方がやや高い結果であった。最適な予後予測モデルはテーマ別に異なると考えられる。近年報告されている新たな生存予測解析モデルも含め、予測テーマ毎に最適モデルを検証することが求められる [7]。

当 Web アプリケーションの利用において、いくつかの制約、注意事項がある。1 つ目は、利便性の観点から予後因子の数が制約されてしまうことである。予後予測曲線を描出する際、患者固有の因子値を設定する必要があるため、数十個以上の因子を予後因子とすることは利便性の面から実用的ではない。ユーザーは、転帰に対してより意義深い予後因子を優先して選択する必要がある。2 つ目は、予後予測モデルに含まれる調整不可能な選択バイアスの存在である。原則として、教師データに含まれないようなシミュレーションにおける予後予測の精度は担保されない。例えば、予後に大きな影響を及ぼすような予防法や治療法が開発され臨床応用され始めた場合、

過去の教師データから作成された予後予測モデルの予測精度を新規患者に当てはめることは適切ではない。ユーザーは教師データに含まれるバイアスを鑑みた上で、提示された予後予測情報を対象患者に当てはめることが妥当かどうかを判断する必要がある。

6. 結論

本 Web アプリケーションを利用することで、臨床医や研究者、各施設・学会は自身が保有する過去の患者データを用いて、診療領域に限定されることなく、新規患者個別の複数の予後因子が考慮された客観的な予測生存曲線を直感的かつ容易に得ることができる。その予測結果は、客観的な情報として患者に提示できると共に、治療選択などの臨床決断の一助にすることができる。

7. 利益相反開示

特記事項なし

8.参考文献

1) Bertsimas D, Dunn J, Pawlowski C1 et al. Applied Informatics Decision Support Tool for Mortality Predictions in Patients With Cancer. JCO Clin Cancer Inform 2018 ; 2 : 1-11.
2) Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random Survival Forests. Ann Appl Stat 2008 ; 2 : 841-60.
3) Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS. High-Dimensional Variable Selection for Survival Data. J Am Stat Assoc. 2010 ; 105 : 205-17.
4) Mogensen, UB,Ishwaran, H,Gerds TA. Evaluating Random Forests for Survival Analysis using Prediction Error Curves. J Stat Softw. 2012 ; 50 : 1-23.
5) Nasejje JB, Mwambi H. BMC Res Notes. Application of random survival forests in understanding the determinants of under-five child mortality in Uganda in the presence of covariates that satisfy the proportional and non-proportional hazards assumption. BMC Res Notes. 2017 ; 10 : 459.
6) Wang H, Li G. A Selective Review on Random Survival Forests for High Dimensional Data. Quant Biosci. 2017 ; 36 : 85-96.
7) Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol. 2018; 18 : 24.