Comparação entre métodos estatísticos e de aprendizado de máquina para predição de risco em síndromes coronarianas agudas

Alleh Nogueira

Introdução

Revisão de Literatura

Objetivo

O objetivo primário deste estudo consiste em comparar o desempenho entre métodos de aprendizado de máquina e de regressão estatística, para predizer mortalidade e reinfarto em pacientes com síndromes coronarianas agudas. Há três objetivos secundários: (1) comparar o desempenho entre os métodos estatísticos e de aprendizado de máquina, para predizer sangramento maior; (2) estimar o *trade-off* teórico entre os riscos cardiovascular e hemorrágico de cada paciente; (3) avaliar as consequências clínicas da aplicação hipotética de cada modelo preditivo.

Métodos

Delineamento e pacientes do estudo

Nosso Registro de Síndromes Coronarianas Agudas (SCA) é uma coorte prospectiva, alocada num hospital soteropolitano terciário. A coorte foi delineada para avaliar os desfechos de pacientes consecutivamente internados em unidade coronariana intensiva, entre setembro de 2011 e julho de 2019. Não houve critérios de exclusão, nem interferência dos pesquisadores na condução do estudo. Todos os participantes proveram consentimento livre e esclarecido por escrito no momento da admissão. O comitê de ética em pesquisa local aprovou o protocolo do estudo sob o certificado de apresentação de apreciação ética .

Incluimos todos os pacientes cujo quadro clínico era compatível com SCA, e que possuiam ao menos um dos seguintes critérios objetivos: (1) injúria miocárdica, definida por variação dos níveis séricos de troponinas cardíacas, com pelo menos um valor acima do 99.º percentil; (2) evidência eletrocardiográfica de isquemia aguda em ao menos duas derivações contíguas, a saber, inversão simétrica de onda T $\geq$ 0,1 mV, infradesnível do segmento ST $\geq$ 0,05 mV, supradesnível de ST $\geq$ 0,1 mV – exceto em V_2 e V_3, onde deve ser $\geq$ 0,25 mV em homens $<$ 40 anos, $\geq$ 0,2 mV em homens $>$ 40 anos, e $\geq$ 0,15 mV em mulheres; (3) obstrução coronariana $\geq$ 70% em portador de síndrome coronariana crônica, e/ou ondas Q patológicas em paciente com história sugestiva de infarto prévio.

Desfechos do estudo

Comparamos regressão estatística a dois modelos de aprendizado de máquina. Na análise primária, cada modelo foi desenvolvido para predizer a ocorrência de um desfecho intra-hospitalar composto de morte por todas as causas e reinfarto não fatal. Na análise secundária, os modelos deveriam predizer a ocorrência intra-hospitalar de sangramento maior, definido como sangramento tipo 3 ou 5 segundo as definições do *Bleeding Academic Research Consortium* (BARC).

Seleção dos potenciais preditores

À admissão, cardiologistas assistentes colheram os dados clínicos, que foram armazenados virtualmente em um conjunto de dados. Com base em conhecimentos prévios, selecionamos 23 variáveis como candidatas a potenciais preditores. Para a regressão logística, a quantidade máxima de potenciais preditores foi determinada pela regra geral de 10 eventos por parâmetro preditor; incluimos como potenciais preditores as variáveis com menor valor de p na análise de associação univariada com o desfecho, dado que fosse estatisticamente significante ($< 0{,}05$). Todas as variáveis candidatas foram incluídas como potenciais preditores nos modelos de aprendizado de máquina, porque não há requisitos formais para o tamanho ideal da amostra.

Características demográficas e da história clínica foram denotadas por 15 variáveis: idade, sexo, índice de massa corporal, uso atual de tabaco, hipertensão arterial sistêmica, diabetes mellitus, dislipidemia, doença arterial obstrutiva periférica, estenose carotídea, doença arterial coronariana (DAC) prévia, infarto agudo do miocárdio prévio, história familiar de DAC prematura (DAC diagnosticada em familiar de 1.º grau com idade < 55 anos em homens e < 65 anos em mulheres), acidente vascular encefálico prévio, doença renal crônica, e uso prévio de aspirina. As 8 demais variáveis referem-se aos exames físico e complementar: frequência cardíaca, pressão arterial sistólica, classes II-IV de Killip, alterações isquêmicas no eletrocardiograma, injúria miocárdica, NT-pró-BNP sérico, creatinina sérica, e hemoglobina sérica.

Derivação e desempenho dos modelos preditivos

A coorte foi dividida aleatoriamente em duas subamostras, segundo uma proporção de 4:1. 80% dos pacientes pertenciam à coorte de derivação, onde se derivou os modelos estatísticos de regressão, e treinou-se os modelos de aprendizado de máquina. Os 20% restantes pertenciam à coorte de validação interna, onde se testou os desempenhos dos modelos. Três modelos foram desenvolvidos para cada desfecho: (1) o modelo estatístico tradicional empregou regressão logística; (2) o modelo básico de aprendizado de máquina empregou o classificador *support vector machine*, que é antigo, simples e amplamente utilizado em predição clínica; (3) o modelo sofisticado de aprendizado de máquina empregou o classificador *extreme gradient boosting*, que é moderno, sofisticado, e pouco estabelecido para predição clínica.

O desempenho de cada modelo foi avaliado de três formas – discriminação, calibração, e acurácia. A discriminação consiste na capacidade do modelo de diferir quem terá o desfecho de quem não o terá. Avaliamos a discriminação pela estatística c de Harrell, que é numericamente igual à área sobre a curva *receiver operator characteristic* (ROC). c é a probabilidade de que qualquer paciente que experimentou o desfecho tenha uma probabilidade predita superior a qualquer um que não experimentou o desfecho. Um c de 1 indica perfeita discriminação; $c \leq 0{,}5$ indica discriminação irrelevante. Comparamos a discriminação do modelo estatístico com a de cada modelo de aprendizado de máquina utilizando os métodos de DeLong.

A calibração denota a concordância entre os desfechos observados e as predições do modelo. Examinamos a discriminação dos modelos mediante inclinação e intercepto das curvas de calibração. A inclinação avalia a dispersão dos riscos estimados e tem um valor alvo de 1. Uma inclinação < 1 sugere extremismo nas estimativas de risco, isto é, o risco dos pacientes em alto risco é superestimado, enquanto o risco dos pacientes em baixo risco é subestimado. Uma inclinação > 1 sugere que as estimativas de risco são bastante conservadoras. O intercepto é estimado pelo *calibration-in-the-large*, e tem um valor alvo de 0; valores negativos sugerem superestimativa do risco, enquanto valores positivos sugerem subestimativa do risco. A acurácia prognóstica dos modelos foi avaliada por sensibilidade, especificidade, valores preditivos e razões de probabilidades; o ponto de corte para classificação dicotômica foi determinado pela estatística J de Youden, onde $J = \text{sensibilidade} + \text{especificidade} - 1$.

Trade-off entre os riscos cardiovascular e hemorrágico

Para cada desfecho – evento cardiovascular adverso maior ou hemorragia maior –, selecionamos o modelo com melhor discriminação. Classificamos os pacientes em decis segundo seus riscos estimados, e calculamos o risco observado de cada decil. Comparamos as classificações de ambos os desfechos pela combinação das categorias de risco estimado. O *trade-off* teórico entre os riscos cardiovascular e hemorrágico foi avaliado pela diferença absoluta, entre o risco cardiovascular observado e o risco hemorrágico observado, de cada decil de risco.

Análise estatística

Variáveis categóricas são descritas por frequências absolutas e relativas e variáveis numéricas por média e desvio padrão ou mediana e intervalo interquartil para distribuições normais e não normais, respectivamente. A normalidade da distribuição foi avaliada por uma combinação de inspeção visual, medidas de assimetria e curtose, e testes formais de normalidade. Utilizou-se o teste t para avaliar diferenças entre variáveis contínuas paramétricas, o teste U de Mann-Whitney para variáveis não paramétricas, o teste χ^2 para variáveis categóricas, e o teste exato de Fisher para tabelas 2x2. Utilizou-se a correção de Bonferroni para duas múltiplas comparações entre as discriminações do modelo estatístico e de cada modelo de aprendizado de máquina. Para tratar as variáveis incompletas, 40 conjuntos de dados multiplamente imputados mediante 100 iterações foram criados e analisados utilizando-se o pacote "mice" [@]. A significância estatística foi definida por um valor de p bicaudal inferior a 0,05. Todas as análises foram conduzidas no R, versão 4.1.0 (*R Foundation for Statistical Computing*, Viena, Austria).

Cálculo do tamanho amostral

Em nossa coorte disponível previamente à concepção deste estudo, observou-se uma incidência de 6,9% do desfecho composto. Utilizando os métodos de Obuchowski [@], estimamos que 1.029 pacientes seriam necessários para detectar uma diferença de 0,12 entre as estatísticas c – com poder de 0,80 e nível de signíficância bicaudal de 0,05 – assumindo-se que ambas as estatísticas c igualem-se a 0,73 sob a hipótese nula e que a razão de alocação seja de 0,069. As premissas embasam-se na estatística c de 0,73 para predição de morte ou reinfarto pelo escore GRACE (*Global Registry of Acute Coronary Events*) [@], e na incidência de 6,9% do desfecho composto nesta coorte, cujos dados estavam previamente disponíveis.

Resultados

Discussão

*Pontos fortes*Para incrementar a representatividade da coorte, não houve critérios de exclusão nem interferência dos pesquisadores para condução do estudo. O número de eventos por variável preditora foi suficiente (o número estimado pelos métodos de Riley foi menor)