(Corradi, Thompson, Mather, Waszynski, & Dicks, 2018)

Corradi, J. P., Thompson, S., Mather, J. F., Waszynski, C. M., & Dicks, R. S. (2018). Prediction of Incident Delirium Using a Random Forest classifier. *Journal of Medical Systems*, *42*(12). https://doi.org/10.1007/s10916-018-1109-0

Abstrato

Delirium é uma complicação médica séria associada a resultados ruins. Dada a complexidade da síndrome, a prevenção e a detecção precoce são fundamentais para mitigar seus efeitos.

Usamos a triagem do Confusion Assessment Method (CAM) e os dados do Electronic Health Record (EHR) para 64.038 visitas de pacientes internados para treinar e testar um modelo de previsão do delirium que surge no hospital. Delirium de incidente foi definido como o primeiro caso de CAM positivo ocorrendo pelo menos 48 horas após o início da internação. Um algoritmo de aprendizado de máquina Random Forest foi usado com dados demográficos, comorbidades, medicamentos, procedimentos e medidas fisiológicas. O conjunto de dados foi particionado aleatoriamente em 80% / 20% para treinamento e validação do modelo preditivo, respectivamente. Dos 51.240 pacientes no conjunto de treinamento, 2.774 (5,4%) experimentaram delirium durante a internação; e dos 12.798 pacientes no conjunto de validação, 701 (5,5%) experimentaram delirium. A subamostragem da população delirium negativa foi usada para resolver o desequilíbrio de classe. O modelo preditivo Random Forest produziu uma área sob a curva de característica de operação do receptor (ROC AUC) de 0,909 (IC de 95% 0,898 a 0,921). Variáveis ​​importantes no modelo incluíram fatores de risco predisponentes e precipitantes previamente identificados. Esta abordagem de aprendizado de máquina exibiu um alto grau de precisão e tem o potencial de fornecer um modelo preditivo clinicamente útil para intervenção precoce em pacientes com maior risco de desenvolver delirium.

**Introdução**

Delirium é uma condição potencialmente letal de alteração do estado mental, atenção e nível de consciência com início agudo e curso flutuante. As taxas relatadas de delirium incidente variaram de 11 a 14% nas enfermarias de medicina geral, 20–29% nas unidades geriátricas e 19–82% nos cuidados intensivos [1]. Vários estudos descobriram que o delirium está associado a resultados ruins (por exemplo, tempo de internação, mortalidade hospitalar, disposição de alta e readmissão), mesmo após o ajuste para fatores adicionais, como idade e gravidade da doença [2-4]. Estudos recentes fornecem evidências de que o delirium também está associado ao declínio cognitivo de longo prazo e a um aumento dos sintomas depressivos [5–8].

Existem vários fatores de risco predisponentes e precipitantes conhecidos para o delirium, mas a fisiopatologia da síndrome ainda é pouco compreendida [9]. Dado o impacto significativo do delirium nos resultados dos pacientes, muita ênfase foi colocada na redução do risco e na detecção precoce [10, 11]. Métodos de teste à beira do leito têm sido desenvolvidos, como o Confusion Assessment Method1 (CAM), que permite a detecção rápida do delirium [12]. No final de 2012, o Hartford Hospital começou a implementar avaliações regulares, uma por turno de enfermagem, de pacientes que usam CAM em áreas de cuidados não críticos e o CAM-ICU em áreas de cuidados críticos [13, 14].

Delirium é uma consequência direta de uma condição médica geral (por exemplo, infecção, insuficiência de órgão), uso ou abstinência de substância intoxicante, exposição a medicamentos ou toxinas ou uma combinação desses fatores. Como tal, a prioridade no tratamento é identificar e tratar a (s) causa (s) subjacente (s) do delirium. As evidências sugerem que a gravidade e/ou duração do episódio de delirium tem alguma relação com desfechos intra-hospitalares insatisfatórios e de longo prazo [15]. Portanto, a detecção precoce é importante para mitigar os efeitos dessa síndrome. A predição precisa do risco individual do paciente permitiria avaliação e intervenção ainda mais cedo no processo patológico.

**Vários estudos têm procurado identificar fatores de risco significativos e / ou produzir modelos para a previsão do delirium [16–19].** Além do foco em populações específicas de pacientes, a maioria dos modelos publicados tem usado amostras pequenas, um conjunto limitado de variáveis ​​preditoras e não levam em conta as mudanças temporais nas medidas clínicas. Nossa hipótese é que o uso de uma abordagem de modelagem flexível e um grande conjunto de dados com muitos recursos produziria previsões precisas. Nós alavancamos um conjunto de dados retrospectivos de avaliações de delirium usando o algoritmo de aprendizado de máquina Random Forest (RF) [20] para gerar um modelo preditivo para delirium incidente em todos os pacientes hospitalizados. O modelo incorporou dados demográficos, comorbidades, procedimentos, medicamentos e medidas que refletem mudanças dinâmicas na fisiologia aguda.

**Material e métodos**

Criação de conjunto de dados

Este estudo retrospectivo foi aprovado pelo conselho de revisão institucional. Durante o período de tempo em que os registros dos pacientes foram analisados, o Hartford Hospital usou o Sunrise Clinical Manager (SCM) (Allscripts, Chicago, IL) como o sistema EHR primário. Para as análises descritas aqui, os registros do paciente com CAMs registrados foram identificados de 1 de setembro de 2012 a 30 de setembro de 2015. A data de término foi escolhida de forma que quase todas as visitas de internação do hospital usassem códigos de procedimento e diagnóstico CID-9. Dados demográficos, comorbidades e medidas fisiológicas cronometradas (sinais vitais), avaliações, pedidos de medicamentos, procedimentos e localização do paciente foram extraídos do EHR usando a Structured Query Language (SQL) para todas as internações. A fim de focar nos casos de delirium incidente, os pacientes "positivos" foram definidos como aqueles com um ou mais CAMs positivos, onde o primeiro positivo foi registrado pelo menos 48 horas a partir do momento da admissão. Pacientes "negativos" foram aqueles com pelo menos 48 horas de permanência para os quais todos os resultados CAM foram negativos. As distribuições de tempo até o evento (primeiro CAM positivo ou alta para pacientes negativos) são mostradas na Figura S1.

**Conjunto de características**

Apenas medições fisiológicas, medicamentos, procedimentos e avaliações ocorrendo antes do primeiro CAM positivo foram considerados para pacientes "positivos"; enquanto os dados de toda a internação hospitalar foram usados para pacientes "negativos". Os dados demográficos básicos incluíram idade, sexo, raça / etnia e estado civil.

Um script Python foi escrito para analisar dados de visitas de pacientes ordenados pelo tempo e calcular estatísticas resumidas para sinais vitais clínicos - pressão arterial, frequência cardíaca, frequência respiratória, temperatura corporal e saturação de oxigênio - com as condições em que havia pelo menos três medidas longitudinais disponíveis e o último valor deve ter sido registrado 8–24 h antes do primeiro resultado CAM positivo.

Além das medidas fisiológicas, a escala Richmond Agitation-Sedation Scale (RASS) estava disponível para a maioria dos pacientes. O RASS é uma avaliação do nível de consciência, com uma escala inteira que varia de −5 (inestimável) a +4 (combativo). O mínimo, máximo e intervalo foram calculados para cada variável fisiológica. A correlação de Pearson entre todas as variáveis fisiológicas é mostrada na Figura S2. Para evitar distorção das estatísticas descritivas para medidas fisiológicas, filtros foram colocados em prática para remover valores claramente errôneos que caíram fora dos extremos da observação clínica.

Apenas visitas de pacientes com mais de 50% dos valores fisiológicos presentes foram usadas nos modelos preditivos. A fração de valores ausentes para cada variável está listada na Tabela S1.

As comorbidades foram contabilizadas pelo cálculo da modificação de Quan do Índice de Comorbidade de Charlson (CCI) [21], usando os códigos de diagnóstico CID-9 e CID-10 e o pacote 'icd' R [22]. Além disso, os códigos da CID também foram usados para identificar as seguintes comorbidades específicas: demência, abuso de álcool e / ou substâncias, transtornos de humor, doenças malignas, fraturas e deficiência visual.

Os procedimentos cirúrgicos, cuidados intensivos, cuidados pós-anestésicos e ventilação mecânica foram capturados para pacientes negativos, enquanto para pacientes positivos apenas se tivessem ocorrido antes do primeiro CAM positivo. As variáveis binárias para cirurgia foram eletivas, emergentes, cardíacas, neurológicas, trauma e cabeça / pescoço.

As ordens de medicação no hospital foram capturadas usando um sinalizador binário (ausente / presente), com a condição para pacientes positivos de terem a primeira ordem pelo menos 8 horas antes do primeiro CAM positivo. Os medicamentos individuais foram mapeados para as categorias disponíveis na Epic Clarity (Epic Systems Corp., Verona, WI), uma vez que o Hartford Hospital migrou para um sistema Epic EHR. A substring correspondente mais longa entre os pedidos de medicamentos SCM e aqueles no Epic Clarity foram mapeados para classes farmacêuticas no Clarity. Os mapeamentos foram inspecionados manualmente e selecionados para reduzir classificações errôneas. Além disso, o número total de medicamentos exclusivos solicitados para cada paciente foi incluído como variável preditora.

Modelagem preditiva

Modelos preditivos foram gerados usando o algoritmo Distributed Random Forest (DRF) implementado na biblioteca de aprendizado de máquina H2O [23], e usado de dentro do ambiente de computação estatística R. Essa implementação lida com valores ausentes tratando-os como informações, enviando-os para uma ramificação com base na minimização de uma função de perda. O algoritmo DRF foi executado com uma semente de randomização usando um núcleo para reprodutibilidade. O conjunto de dados foi dividido aleatoriamente em um conjunto de treinamento contendo 80% das amostras e um conjunto de validação contendo os 20% restantes. Uma pesquisa de grade foi empregada nos dados de treino com validação cruzada para ajustar os parâmetros de RF para tamanho de seleção de variável (mtries = √p, onde p é o número de variáveis ​​preditoras), número de árvores a gerar (ntrees = 200), e profundidade máxima da árvore (max\_depth = 30). Para abordar o grau de desequilíbrio nas classes (positivos foram 5,5% do conjunto de amostra), a maioria das classes (negativas) foi sub-amostrada em cada etapa de bootstrapping [24, 25]. Especificamente, uma amostra bootstrap da classe negativa equivalente em tamanho à classe positiva foi selecionada para a construção da árvore. Para derivar probabilidades de classe mais precisas do modelo Random Forest, a calibração por escalonamento de Platt (uma transformação logística dos escores de classificação [26]) foi implementada dentro do modelo H2O usando o conjunto de dados de validação. Um gráfico de calibração foi gerado para comparar as pontuações da classificação original com as pontuações calibradas usando o 'acento circunflexo' do pacote R [27] (Figura S3).

Todos os valores ROC AUC relatados e as métricas relacionadas à matriz de confusão foram calculados em previsões usando os dados de validação e pontuações de classificação calibradas. O intervalo de confiança de 95% do ROC AUC do conjunto de validação foi calculado com 2.000 réplicas estratificadas bootstrapped usando o pacote R 'pROC' [28]. O pacote R 'PRROC' foi usado para gerar o ROC e as curvas de recall de precisão (PR) e a área sob os cálculos da curva [29]. Exatidão, precisão, recall e especificidade foram relatados em pontos operacionais escolhidos para maximizar as métricas de desempenho comumente usadas. A precisão média por classe é simplesmente a média de recall e especificidade. As 'medidas F' são meios harmônicos ponderados de precisão e recall. Eles são calculados usando esta fórmula:

O coeficiente de correlação de Matthews (MCC) é uma medida balanceada considerando verdadeiros positivos (TP), verdadeiros negativos (TN), falsos positivos (FP) e falsos negativos (FN). É definido como:

O algoritmo H2O DRF pontua a importância da variável pela melhoria no erro quadrado quando uma variável é selecionada durante uma divisão, calculada a média de todas as árvores. As listas de variáveis relatadas aqui são mostradas classificadas pela importância relativa escalonada para a variável mais importante, com um limite mínimo de 0,05.

O algoritmo H2O DRF pontua a importância da variável pela melhoria no erro quadrático quando uma variável é selecionada durante uma divisão, calculada a média de todas as árvores. As listas de variáveis ​​relatadas aqui são mostradas classificadas pela importância relativa escalonada para a variável mais importante, com um limite mínimo de 0,05. Resultados Os pacientes do Hospital Hartford são avaliados para delirium usando o CAM durante cada turno de oito horas e com uma mudança na condição. Para o período de 37 meses, de setembro de 2012 a setembro de 2015, houve 96.302 visitas de pacientes internados de adultos compostas por 58.973 pacientes únicos no registro de delirium do Hospital Hartford. Dessas visitas, 7.613 (7,9%) tiveram pelo menos um CAM positivo registrado. A fim de focar apenas no delírio incidente para modelagem preditiva, a população do estudo foi restrita aos pacientes cujo primeiro CAM positivo ocorreu pelo menos 48 h em sua permanência (o grupo 'Positivo'), bem como aqueles pacientes com apenas registros negativos do CAM (ou seja, sem resultados 'positivos' ou 'incapazes de avaliar') que estiveram no hospital por pelo menos 48 horas (o grupo 'Negativo'). Esta seleção rendeu 64.237 visitas de pacientes, 3.499 (5,4%) das quais estavam no grupo Positivo. Uma restrição adicional colocada na coorte foi a disponibilidade de dados fisiológicos longitudinais. Após excluir os pacientes com dados de sinais vitais insuficientes, o conjunto restante consistiu em 64.038 visitas de pacientes (41.826 pacientes únicos), 3499 (5,5%) das quais estavam no grupo Positivo (fig. 1).

Um modelo Random Forest (RF) foi treinado em 80% dos dados da visita do paciente usando todas as 128 variáveis que abrangem dados demográficos, sinais fisiológicos, comorbidades, procedimentos e medicamentos (Tabela S1). O desempenho do modelo foi avaliado em 20% do conjunto de validação apresentado (ver Tabela 1

para composição de conjuntos de treinamento e teste). O modelo foi capaz de prever o delírio incidente no conjunto de validação com uma área sob a curva de característica de operação do receptor (ROC AUC) de 0,909 (IC de 95% 0,898 a 0,921) (fig. 2). As taxas de erro de treinamento e teste foram comparáveis, indicando que o modelo não superaqueceu os dados de treinamento. Para a previsão de classes não balanceadas, também é útil inspecionar a curva de recall de precisão (PR). O PR AUC para o modelo foi de 0,604 (Fig. 2), e no escore F2 máximo - uma média ponderada de precisão e recall que favorece o recall - a precisão foi de 0,357, o recall (sensibilidade) foi de 0,698 e a especificidade foi de 0,927 ( Mesa 2).

Os modelos de floresta aleatória podem fornecer medidas de importância variável, permitindo alguns insights sobre os fatores que mais influenciam as previsões. As variáveis altamente classificadas neste modelo foram nível de consciência (conforme medido pelo RASS) e alterações de sinais vitais, cuidados intensivos, ventilação mecânica e outros fatores de risco predisponentes conhecidos, como idade, demência, carga de comorbidade e

polifarmácia (Fig. 3).

Após a inspeção dos resultados da predição, ficou evidente que o modelo geral teve um desempenho um pouco melhor para os pacientes da UTI. Para fornecer uma visão mais detalhada da utilidade de um modelo de RF para delirium em diferentes pacientes

populações, o conjunto de amostras foi dividido em subgrupos de cuidados intensivos e de cuidados não críticos. O conjunto de amostra de cuidados intensivos continha 13.676 visitas de pacientes, 2.388 (17,5%) das quais foram positivas para delirium. Houve 50.362 consultas de cuidados não críticos com 1.087 (2,2%) no grupo Positivo. Novos modelos de RF foram gerados para cada coorte usando o mesmo processo descrito para o modelo completo. Um modelo treinado em pacientes de UTI atingiu uma AUC de ROC de 0,930 (IC de 95% 0,918 a 0,943), enquanto um modelo treinado em pacientes não UTI rendeu uma AUC de ROC de 0,861 (IC de 95% 0,839 a 0,884) (Tabela 3).

Variáveis importantes na população de UTI foram dominadas por alterações fisiológicas agudas, especialmente no nível de consciência (RASS), enquanto fatores predisponentes como idade, demência e comorbidades figuraram com mais destaque na previsão de delirium fora dos cuidados intensivos (Fig. 4). Um número maior de variáveis atingiu o mesmo limite de importância nas enfermarias médicas gerais.

Discussão

Este estudo de prova de conceito demonstra o valor do aprendizado de máquina aplicado a grandes e complexos conjuntos de dados clínicos.

A predição precisa e geralmente aplicável de pacientes com alto risco de desenvolver delirium seria de grande valor na identificação de condições médicas subjacentes e fatores de risco modificáveis. A redução das taxas de delirium hospitalar e do tempo gasto no delirium melhoraria os resultados para os pacientes, bem como diminuiria a carga geral de cuidados e os custos associados para o sistema de saúde. Os fatores incorporados neste modelo podem ser preditivos, mesmo que não sejam precipitantes (por exemplo, mudanças nos sinais vitais ou nível de consciência.

Infelizmente, nem todos os fatores precipitantes estão universalmente disponíveis para os pacientes, quando seriam preditivos. Muitas vezes, o delírio incidente é a síndrome clínica que leva à busca da causa subjacente, como infecção, disfunção orgânica, toxicidade, etc. O modelo preditivo descrito aqui pode ser usado para iniciar essas ações, em vez de esperar até que haja delírio franco.

A precisão geral alcançada pelo modelo de RF descrito aqui é comparável ou excede a de outros modelos preditivos publicados para delirium, apesar de ser aplicada a uma população diversificada de pacientes. Dois dos modelos relatados de melhor desempenho, um para pacientes de cirurgia geral e um para pacientes de cuidados intensivos, usaram um número maior de variáveis preditoras ou pacientes do que a maioria dos outros estudos [17, 18]. Nosso modelo incorpora um conjunto muito maior de recursos e dezenas de milhares de visitas de pacientes.

Embora algumas dessas variáveis possam não ter sido relevantes para o modelo final, há vários motivos para sua inclusão. Primeiro, as Florestas Aleatórias geralmente não são afetadas pela adição de variáveis irrelevantes, e são um tanto robustas à colinearidade devido ao uso de subconjuntos de variáveis aleatórias para divisões de árvores [30]. Em segundo lugar, todos os dados incluídos neste modelo estavam prontamente disponíveis no EHR. Finalmente, dada a complexa etiologia multifatorial do delirium, pode haver preditores inesperados ou subestimados. Dito isso, um modelo simplificado pode ser gerado usando um conjunto de variáveis reduzido com base na importância da variável. Na verdade, o uso das 48 principais variáveis preditoras classificadas por importância relativa (Tabela S1) produziu um modelo com desempenho comparável (não mostrado).

O programa de avaliação de delirium do Hartford Hospital (‘ADAPT’ [31]) possui dois pontos fortes críticos que facilitaram este esforço de modelagem preditiva. Ter o CAM realizado três vezes ao dia fornece a resolução temporal necessária para detectar mudanças na condição de um paciente na escala de horas. Delirium é uma síndrome dinâmica, e os fatores precipitantes que contribuem para o risco estão mudando na mesma escala de tempo. Além disso, o RASS fornece uma medida independente do nível de excitação, ao mesmo tempo que é usado como um componente do CAM. Nosso modelo sugere que o RASS longitudinal é um indicador sensível do risco de delirium, e essa observação é consistente com estudos publicados [32-34].

As abordagens de modelagem preditiva oferecem diferentes pontos fortes, com uma compensação entre complexidade e interpretabilidade. No caso de previsão médica, especialmente de uma condição urgente associada a resultados ruins, a precisão deve ser primordial. No entanto, no ambiente clínico, a confiança nas ferramentas de diagnóstico ou prognóstico também é importante. Por esse motivo, Random Forests fornecem um bom equilíbrio entre precisão e compreensão. Embora não forneçam regras simples ou categorias de risco, eles permitem "dar uma olhada nos bastidores" para ver quais variáveis que influenciam a previsão. Se os fatores importantes identificados forem consistentes com o conhecimento clínico e a intuição, haverá mais confiança no modelo como ferramenta de suporte à decisão. Além disso, Random Forests pode acomodar muitas variáveis contínuas e categóricas sem a necessidade de transformação ou seleção de variáveis, pode modelar interações complexas e ter poucos parâmetros para ajustar.

Por ser um estudo retrospectivo, houve limitações impostas pela disponibilidade de dados estruturados para inclusão. Uma omissão saliente foi a dos valores dos testes laboratoriais, especialmente dada a identificação de medidas específicas associadas ao risco de delirium (por exemplo, razão BUN / creatinina, eletrólitos). Embora a maioria dos pacientes internados tivesse valores de teste disponíveis, muitos não tinham dados longitudinais suficientes antes do início do delirium para capturar as mudanças dinâmicas que podem preceder a condição. Descobrimos que a exclusão dos valores dos testes laboratoriais aumentou significativamente o tamanho da amostra do paciente disponível para modelagem e não houve perda subsequente na precisão da previsão.

Uma limitação adicional do uso de dados disponíveis é a confiabilidade da medida de resultado - ou seja, as avaliações de delirium.

Os resultados do CAM usados ​​neste estudo foram realizados por muitos membros do corpo clínico no atendimento de rotina de pacientes hospitalares, sem validação independente. Como tal, haverá algumas avaliações positivas que podem ter sido feitas por engano; devido a sintomas confusos de demência ou psicoses, ou a uma correspondência incompleta com os critérios para um CAM positivo. Por outro lado, haverá, sem dúvida, delírio não reconhecido entre o grupo CAM negativo. Um resultado do CAM é ‘não é possível avaliar’ (UTA), uma classificação que deve ser reservada para os casos em que o paciente não é capaz de reconhecer a presença do examinador. No entanto, encontramos anteriormente casos do uso de UTA fora de cuidados intensivos, sugerindo que pode mascarar casos de delirium, com base em avaliações e resultados adicionais do paciente [35]. Para reduzir o impacto desses resultados ambíguos, confinamos nosso grupo de delirium negativo aos pacientes que tiveram resultados CAM exclusivamente negativos (ou seja, nenhum CAM UTA).pacientes que tiveram resultados CAM exclusivamente negativos (ou seja, nenhum CAM UTA).

Aplicações recentes de aprendizado de máquina (incluindo, mas não se limitando a Florestas Aleatórias) para dados médicos têm se mostrado uma grande promessa, de diagnósticos por imagem [36, 37] à previsão de resultados ruins associados a uma variedade de condições [38-42]. Durante a preparação deste relatório, Kramer et al. publicou uma comparação preliminar de modelos multivariados para prever delirium em uma grande população de pacientes onde Random Forest teve um bom desempenho [43]; no entanto, não havia dados suficientes disponíveis para uma comparação com nosso modelo. Dois estudos recentes adicionais sobre delirium empregando aprendizado de máquina foram publicados enquanto este manuscrito estava sendo revisado. Wong et al. comparou vários algoritmos (incluindo Random Forest) para prever o risco de delírio incidente usando apenas as informações disponíveis nas primeiras 24 horas de admissão [44]. Ao contrário do nosso modelo, os autores excluíram populações de alto risco em cuidados intensivos ou com comprometimento cognitivo conhecido. Acreditamos que nosso modelo permite uma predição mais dinâmica, usando informações clínicas atualizadas, e que a predição de risco de paciente individual ainda é muito valiosa, mesmo em casos de alto risco conhecido. populações. No estudo de Halladay et al., Os autores usaram Random Forest para identificar os fatores mais importantes úteis como um auxílio diagnóstico no reconhecimento de delirium prevalente, ao invés de um modelo preditivo para delirium incidente [45].

A nossa abordagem de aprendizado de máquina, embora exiba excelente validação interna com o maior coorte de treino que conhecemos, exigirá validação prospectiva para ser considerada para uma implementação em tempo real. As Florestas Aleatórias não produzem sistemas de pontuação simples ou regras de decisão clínica; e, portanto, esperamos que seja a abordagem em si que será generalizável. Os modelos de aprendizado de máquina podem ser treinados e otimizados usando dados locais, levando em consideração as diferenças nas medidas de avaliação, formatos de relatórios e protocolos institucionais. Isso geralmente requer mais automação e processamento para funcionar como ferramentas eficazes de suporte à decisão. No entanto, a precisão proporcionada por tais modelos poderia justificar o esforço adicional de implementação.