Validação de Escores Prognósticos para Mortalidade de Pacientes com COVID-19 no Brasil

Lucas Emanuel Ferreira Ramos

Departamento de Estatística
Universidade Federal de Minas Gerais

Belo Horizonte
2021

Sumário

LUCAS EMANUEL FERREIRA RAMOS

VALIDAÇÃO DE ESCORES PROGNÓSTICOS PARA MORTALIDADE DE PACIENTES COM COVID-19 NO BRASIL

Projeto de Monografia sob Orientação da Profa. Dra. Magda Carvalho Pires (DEST/UFMG) e Co-orientação da Profa. Dra. Milena Soriano Marcolino (Departamento de Clínica Médica/UFMG)

Validação de Escores Prognósticos para Mortalidade de Pacientes com COVID-19 no Brasil

Escores prognósticos são desenvolvidos para quantificar a gravidade de doenças e direcionar as intervenções clínicas e terapêuticas. Esses escores estratificam os pacientes em níveis de risco para guiar o gerenciamento dos centros médicos e otimizar a alocação de recursos. Em um cenário pandêmico, como a pandemia causada pelo coronavírus e declarada pela Organização Mundial da Saúde (OMS) em março de 2020, um índice prognóstico que permita a precoce identificação de pacientes com COVID-19 os quais apresentam maiores riscos de óbito seria uma ferramenta de extrema urgência e importância. Para isso, a devida validação se apresenta como uma etapa crucial para garantir a acurácia preditiva e minimizar a incerteza na predição de novos pacientes. Neste trabalho, são apresentadas técnicas e métodos para avaliar a discriminação, calibração, desempenho geral e utilidade clínica de escores prognósticos e suas aplicações no escore $ABC_2 - SPH$ de estratificação de risco desenvolvido para predizer o risco de óbito de pacientes admitidos em hospitais públicos e privados no Brasil com COVID-19. Ao final das análises foi possível validar a utilização do escore $ABC_2 - SPH$ em pacientes com COVID-19 no Brasil e estudar sua aplicação em pacientes da Espanha com COVID-19 através de uma validação externa.

Palavras-chave: COVID-19; estratificação de risco; validação; calibração; discriminação; escore.

Aluno: Lucas Emanuel Ferreira Ramos
Orientadora: Profa. Magda Carvalho Pires (DEST/UFMG)
Co-orientadora: Profa. Milena Soriano Marcolino (Departamento de Clínica Médica/UFMG)

Belo Horizonte
Março/2021

1 Introdução

Ferramentas de predição de risco têm sido cada vez mais utilizadas no meio clínico para estimar e estratificar o risco de pacientes desenvolverem doenças ou mesmo de irem a óbito devido a complicações causadas por essas doenças. Com isso, cresce o interesse em métodos usados para validação. A validação de ferramentas de predição de risco é necessária para determinar a generalização das mesmas, permitindo o seu uso e garantindo boa eficácia em outras populações.

As duas principais medidas usadas para avaliar o desempenho de uma ferramenta de predição de risco são a calibração e discriminação. A calibração está relacionada com a concordância entre os desfechos observados e preditos, enquanto a discriminação é a capacidade de classificar com precisão os indivíduos de baixo e alto risco. Além disso, a utilidade clínica é importante para medir a habilidade do modelo em fazer classificações melhor do que os métodos clínicos padrão (sem a predição do modelo).

Dessa forma, um bom escore de estratificação de risco deve permitir a classificação razoavelmente confiável de pacientes em grupos de risco com prognósticos diferentes. E, para demonstrar essa capacidade, não é suficiente avaliar apenas o seu desempenho nos dados utilizados no desenvolvimento da própria ferramenta. É necessário estender a avaliação para se ter evidências de que o modelo funciona bem para outros grupos de pacientes.

Este trabalho tem o objetivo de apresentar as técnicas que podem ser utilizadas para validar escores prognósticos com foco em uma ferramenta de predição de risco, denominada $ABC_2 - SPH$ (Marcolino et al., 2021), desenvolvida para estratificar o risco de óbito de pacientes admitidos em hospitais públicos e privados do Brasil com COVID-19. Além de avaliar a aplicação deste escore em pacientes brasileiros, será estudada seu desempenho em uma amostra de pacientes espanhóis infectados pelo SARS-CoV-2.

2 Metodologia

2.1 $ABC_2 - SPH$: Um Escore de Estratificação de Risco para Pacientes com COVID-19

Antes de explorar as técnicas e métodos de validação é importante conhecer o caso em estudo. Para a construção do escore foram coletadas informações de pacientes admitidos em diversos hospitais públicos e privados do Brasil com COVID-19 entre março e setembro de 2020 a partir do sistema REDCap (Harris et al., 2009). Os dados foram coletados, tratados e inconsistências foram verificadas e corrigidas. A base de dados, entretanto, possui dados faltantes, um problema muito comum em bases de dados clínicos. Para evitar a perda de informação e um possível viés por exclusão dos casos incompletos, o banco de dados foi submetido à imputação múltipla por equações em cadeia (MICE) (Azur et al., 2011). Dessa forma, foram obtidos $m = 10$ bancos imputados. Seguindo as recomendações (Marshall et al., 2009), todos os bancos de dados imputados foram submetidos às técnicas estatísticas de construção do escore e os resultados combinados utilizando as *Regras de Rubin* (Rubin, 1988).

Os pacientes foram divididos em base de derivação e validação de acordo com a data de admissão. Variáveis foram pré-selecionadas de acordo com a disponibilidade na admissão do paciente no centro médico e por evidências na literatura pela relação com o agravamento da doença causada pelo coronavírus e outras doenças respiratórias. Variáveis com mais de 33% de dados faltantes foram desconsideradas. Das variáveis restantes foi feita uma seleção utilizando Modelos Aditivos Generalizados (GAM). As variáveis contínuas do modelo final foram categorizadas a fim de simplificar o cálculo do escore e as mesmas foram introduzidas em modelos de regularização LASSO juntamente com as demais variáveis categóricas permanecentes. Os coeficientes (pesos) da regressão LASSO foram escalados conforme necessário para determinar o escore prognóstico em uma escala conveniente. Dessa forma, utilizando de variáveis disponíveis na admissão do paciente é possível calcular seu escore que quantificará o risco de óbito do mesmo. O escore assume valores entre 0 e 20 e foi agrupado em 4 grupos de risco (Baixo: 0-1; Intermediário: 2-4; Alto: 5-8; e Muito Alto: 9-20) de acordo com a probabilidade predita pelo modelo prognóstico: $< 6\%$; $6 - 14,9\%$; $15 - 49,9\%$ e $\geq 50\%$, respectivamente.

É importante ressaltar que, no caso em estudo, como foi utilizada a imputação múltipla, existe mais de um banco de dados que possa ser utilizado para validação do escore, uma vez que tanto a base de derivação do escore quanto a base de validação foram imputadas.

2.2 Tipos de Validação

A performance do modelo pode ser avaliada usando novos dados da mesma fonte que a amostra de derivação, mas uma verdadeira avaliação de generalização requer a avaliação com dados de outra fonte. É possível ordenar as estratégias de validação de acordo com o nível de rigor da seguinte forma: (Altman et al., 2009)

1. *Validação Interna*: Uma abordagem comum é dividir o banco de dados aleatoriamente em duas partes. O modelo é desenvolvido utilizando a primeira parte (também chamada de base de treino), e a segunda porção é utilizada para avaliar a acurácia preditiva do modelo. Essa abordagem tende a dar resultados otimistas, pois as duas bases são muito similares. Essa estratégia é útil, mas não fornece informações acerca do desempenho do modelo em outros dados.

2. *Validação Temporal*: Uma alternativa para avaliar o desempenho do modelo em pacientes subsequentes dos mesmos centros médicos é simplesmente particionar o banco de dados pelo tempo. Claramente haverá muitas similaridades entre os dois conjuntos de pacientes, entretanto, a validação temporal é uma avaliação prospectiva do modelo, independente dos dados originais usados no processo de desenvolvimento. Essa estratégia pode ser considerada externa no tempo e, portanto, intermediária entre validação interna e externa.

3. *Validação Externa*: Para examinar a generalização do modelo é necessário usar novos dados coletados de uma população de pacientes apropriada e similar em diferentes centros.

Para o caso em estudo, a partição da base de dados foi feita por meio da data de admissão dos pacientes, permitindo uma validação temporal que garante certo rigor aos resultados. Para estudar, de fato, a generalização do modelo é necessário obter novos dados de outros pacientes de centros médicos similares ou utilizar dados retrospectivos.

2.3 Medidas de Desempenho Geral

A distância entre o desfecho predito $\hat{y}$ e o desfecho observado y é uma maneira de quantificar de modo geral a performance do modelo de uma perspectiva estatística (Steyerberg et al., 2019). Essa distância, também chamada de resíduo, é definida como $y - \hat{y}$ para respostas contínuas. Quando o desfecho é binário, como o caso em estudo (óbito ou não-óbito), $\hat{y}$ é igual à probabilidade predita p. Bons modelos têm menores distâncias entre os desfechos preditos e observados e estão relacionadas a uma melhor bondade ou qualidade de ajuste (*goodness-of-fit*) e a uma maior variabilidade explicada pelo modelo.

2.3.1 Escore de Brier (*Brier Score*)

O escore de Brier é muito utilizado para avaliar a acurácia de modelos preditivos binários em estudos clínicos e é uma alternativa à variação explicada (R^2) calculada para desfechos contínuos. Seu valor é definido como:

$$BS = \frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2, \tag{2.1}$$

em que y_i é o desfecho observado e p_i é a probabilidade predita pelo modelo. Dessa forma, o escore varia de 0 a 1, de modo que quanto mais próximo de zero, menor a distância média entre predito e observado e, assim, melhor a qualidade do ajuste.

2.3.2 Decomposição em Discriminação e Calibração

Medidas de performance geral incorporam tanto aspectos de discriminação quanto de calibração. A discriminação está relacionada a quão bem um modelo preditivo consegue diferenciar os pacientes com diferentes desfechos. Já a calibração está relacionada com a concordância entre os desfechos observados e preditos. O estudo da calibração e discriminação do modelo em estudo é mais relevante, então, do que uma medida geral como o escore de Brier.

2.4 Discriminação

Modelos de predição para resposta binária precisam discriminar entre os pacientes que possuem e não possuem o evento de interesse. Diversas medidas podem ser utilizadas para indicar quão bem um modelo classifica pacientes em problemas binários. A área abaixo da curva ROC (*receiver operating characteristic*), denominada de AUC, é a mais comumente usada para descrever a habilidade de discriminação de modelos preditivos binários. A curva ROC traça a sensibilidade (taxa de verdadeiros-positivos) *versus* a taxa de falsos-positivos (1 - especificidade) para diferentes pontos de corte de classificação para a probabilidade predita pelo modelo. Entretanto, é comum se traçar a sensibilidade *versus* a especificidade invertendo o eixo x como na Figura 2.1.

A sensibilidade é definida como a proporção dos pacientes que foram corretamente classificados pelo modelo dentre aqueles que possuem o evento de interesse, dado um ponto de corte para a probabilidade predita. Já a especificidade é a proporção dos pacientes que também foram corretamente classificados, mas que não possuem o evento de interesse. Sendo assim, 1 - especificidade representa o percentual de pacientes que não foram classificados corretamente dentre aqueles que não possuem o evento. Dessa forma, o interesse está em obter alta sensibilidade (taxa de verdadeitos-positivos) e alta especificidade e, consequentemente, baixa taxa de falsos-positivos.

2.4.1 Curva ROC e AUC

A curva ROC é traçada variando o ponto de corte nas probabilidades preditas para definir, assim, a classificação. O ponto de corte inicial é 0% fazendo com que todos os pacientes sejam classificados como positivos (presença do evento), resultando em uma sensibilidade de 100% e uma especificidade de 0%. O ponto de corte varia até chegar a 100%, em que a sensibilidade é 0% e a especificidade é 100%.

A área abaixo da curva ROC (AUC) pode ser interpretada como a probabilidade de que um paciente com o evento de interesse tenha uma maior probabilidade atribuída pelo modelo do que um paciente escolhido aleatoriamente que não tenha o evento (Hanley and McNeil, 1982). Dessa forma, como o valor da área varia entre 0 e 1 e um modelo não informativo terá uma área de 0.5, quanto mais próximo de 1, maior a discriminação do modelo preditivo.

A Figura 2.1 mostra diferentes curvas ROC e, consequentemente, diferentes áreas abaixo da curva. A curva em preto representa o modelo não informativo e possui uma AUC de 0.5, enquanto que a curva em azul representa o modelo perfeito com área 1. O ideal é obter um modelo prognóstico cuja curva ROC se aproxime à curva azul.

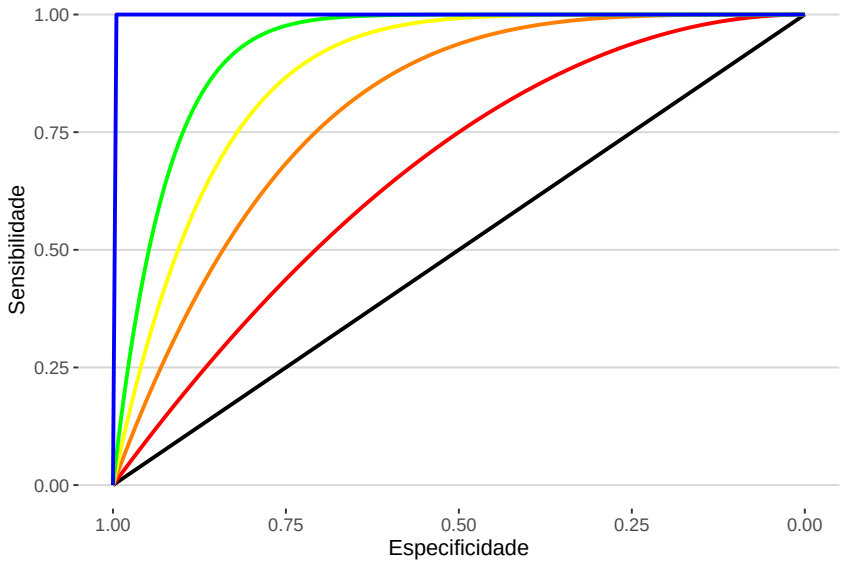

Figura 2.1: Exemplo de 6 diferentes curvas ROC. No eixo x foi utilizada a especificidade ao invés de 1 - especificidade com o eixo invertido para ser mantido o mesmo comportamento padrão da curva.

Importante ressaltar que a AUC é considerada insensível a melhorias na predição ao adicionar preditores. Por outro lado, melhorias no ajuste do modelo estão diretamente relacionadas com melhorias na AUC, se o modelo corrente estiver correto. Dessa mesma forma, outros indicadores como Escore de Brier e R^2 devem mostrar melhorias se um preditor verdadeiro for adicionado ao modelo.

Intervalos de confiança para a AUC podem ser calculados a partir de vários métodos. Reamostragem *bootstrap* é uma boa alternativa em diversas situações.

2.5 Calibração

Outra propriedade crucial na predição de um modelo é a calibração, isto é, a concordância entre valores observados e preditos. A calibração geralmente é entendida da seguinte forma: se for observado um risco $p\%$ em um grupo de pacientes, é esperado que seja predito um risco próximo de $p\%$ pelo modelo - esta é a chamada "calibração moderada". Formas mais "fracas" de calibração requerem apenas o risco médio predito (calibração média) ou os efeitos de predição médios (calibração fraca). A denominada "calibração forte" requer que a taxa de eventos seja igual a taxa predita para cada padrão de covariável. A "calibração forte" é desejada, mas irrealista, além de estimular o desenvolvimento de modelos excessivamente complexos. Portanto, é recomendado que o desenvolvimento e validação do modelo se concentrem na "calibração moderada" (Van Calster et al., 2016).

2.5.1 Gráfico de Calibração (*Calibration Plot*)

O gráfico de calibração possui as probabilidades preditas no eixo x e as probabilidades observadas no eixo y. Em modelos binários são observados desfechos 0/1 e, com isso, as probabilidades não são observadas diretamente. Geralmente utiliza-se de funções de suavização para estimar as probabilidades observadas em relação às probabilidades preditas.

As respostas binárias são substituídas por valores entre 0 e 1 ao combinar os desfechos de pacientes com probabilidades preditas similares. Podem ser utilizados algoritmos como *loess* e *lowess* resultando em uma curva não paramétrica. Também pode ser utilizada a regressão logística como na equação (2.2).

A "calibração média" (*calibration-in-the-large*) é obtida pela diferença entre o risco médio predito e a taxa de incidência do evento observada. Entretanto, existem outros tipos de calibração mais informativos.

A Figura 2.2 mostra uma curva de calibração em que os dados são separados em 5 grupos e são traçadas curvas de calibração logística e não paramétrica (*lowess*). O ideal é que as duas curvas se aproximem da reta identidade (45º) indicando que as probabilidades preditas são muito próximas às observadas. No gráfico é possível visualizar a "calibração fraca" (existe uma tendência geral de superestimação ou subestimação do risco ou das previsões serem muito extremas, um sinal de ajuste excessivo?). Além disso, agrupando os dados em grupos com probabilidades similares, é possível avaliar a "calibração moderada" ao comparar a proporção média observada e a média das probabilidades preditas por grupo. É comum formar grupos por quantís de probabilidade. (Van Calster et al., 2019) É possível, ainda, ver um pouco da discriminação no gráfico. Um modelo com maior discriminação terá os pontos correspondentes a cada agrupamento mais separados. Claro que a escolha dos grupos é importante para a visualização: grupos menores apresentarão maior variabilidade.

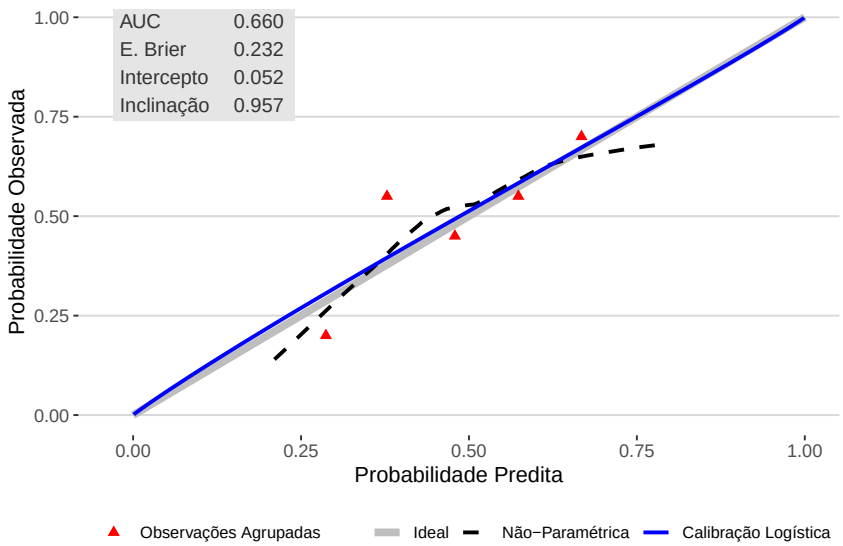

Figura 2.2: Exemplo de um gráfico de calibração com dados simulados e grupos de 20 observações. A linha cinza é a reta identidade e a linha azul representa a calibração logística. Já a linha pontilhada mostra uma curva não paramétrica de calibração.

2.5.2 Intercepto e Inclinação de Cox

No canto superior esquerdo da Figura 2.2 são exibidas algumas estatísticas que podem auxiliar na avaliação do modelo preditivo. Entre elas estão a área abaixo da curva ROC (AUC) e o escore de Brier (E. Brier). Outros índices que são amplamente usados na avaliação de escores prognósticos são o intercepto e inclinação de Cox que medem a "calibração fraca". O método de Cox avalia a calibração ao ajustar uma regressão logística que explica

o desfecho binário observado ($O = 1$) por meio das *log odds* das estimativas do modelo preditivo em estudo (E):

$$logit\{P(O = 1)\} = a + b * logit(E), \tag{2.2}$$

em que b é a inclinação da regressão e a é o intercepto. A estimativa da inclinação indica a direção da descalibração, de modo que 1 denota a calibração perfeita, > 1 denota subestimação do alto risco e superestimação do baixo risco, e < 1 denota subestimação do baixo risco e superestimação do alto risco. O intercepto indica a descalibração geral, em que 0 significa boa calibração, > 0 denota subestimação média, e < 0 denota superestimação média (Huang et al., 2020). Dessa forma, é esperado que um modelo com boa calibração tenha o intercepto próximo de 0 e a inclinação próxima de 1.

É possível realizar um teste de hipóteses para testar $H_0 : a = 0, b = 1$. O teste qui-quadrado com 2 graus de liberdade compara os desvios dos modelos encaixados (2.2) e:

$$logit\{P(O = 1)\} = offset\{logit(E)\}. \tag{2.3}$$

2.6 Avaliação da Utilidade Clínica

Além de avaliar a performance do modelo por meio da discriminação e calibração, também pode ser interessante avaliar o quanto um modelo preditivo é útil para auxiliar na tomada de decisão clínica: o modelo é benéfico para guiar a seleção de pacientes para triagem, investigação diagnóstica ou terapia? Para isso, é necessário um ponto de corte para as probabilidades preditas para definir a classificação. O termo "utilidade clínica" é usado para a habilidade do modelo em fazer classificações melhor do que os métodos padrão (sem a predição do modelo). Isto significa considerar medidas de performance para a classificação em uma perspectiva de decisão analítica.

No caso binário um ponto crucial é a escolha do ponto de corte de classificação dos pacientes. Nos *softwares* estatísticos, por padrão, o corte é 50%, ou seja, pacientes com probabilidade predita acima de 50% são classificados como positivos (possui o evento de interesse). Isso implica que as taxas de falso-positivo e falso-negativo são igualmente importantes, o que pode ser considerado incorreto em problemas médicos de predição. A perda de um paciente com o evento é usualmente mais importante do que uma incorreta classificação de um paciente sem o evento. Erros de falso-negativo são mais importantes do que falso-positivo. Isso implica em um ponto de corte menor do que 50%. O ponto de corte ótimo é definido em um contexto de decisão, não por critério estatístico. Uma vez que o ponto de corte é definido, medidas de utilidade clínica podem ser definidas.

2.6.1 Análise da Curva de Decisão (*Decision Curve Analysis*)

Na prática, há uma dificuldade em definir o ponto de corte ótimo precisamente. É possível obter alguma noção do ponto de corte típico através de especialistas clínicos. Uma abordagem atrativa é estudar uma gama de pontos de corte plausíveis e, com isso, construir um gráfico que mostra o que é chamado de "benefício líquido" (*net benefit*) de tratar pacientes

de acordo com o modelo de predição. Esta medida pode ser obtida por meio da fórmula: (Steyerberg et al., 2019)

$$NetBenefit = NB = (TP - w * FP)/n, \tag{2.4}$$

em que TP é o número de classificações verdadeiro-positivo, FP o número de falso-positivo, w o peso definido como a *odds* do respectivo ponto de corte $\frac{p_t}{(1-p_t)}$, e n é o número total de pacientes. O ponto crucial é o peso w. Por exemplo, um ponto de corte de 10% significa um peso $w = 1/9$: as classificações falso-positivo são avaliadas em um nono de classificações verdadeiro-positivo, fazendo com que o peso da classificação incorreta de um paciente com o evento seja 9 vezes maior do que de um paciente sem o evento. O benefício líquido de um modelo preditivo deve ser comparado com as políticas padrão "não tratar ninguém" e "tratar todos". Ao não tratar ninguém, o NB é zero. Ao tratar todos, o NB depende da taxa de incidência do evento: o NB é positivo até o ponto em que o ponto de corte atinge a taxa de incidência.

A Figura 2.3 mostra um exemplo de curva de decisão em dados simulados com incidência de 50% em que o modelo em análise apresenta pouco benefício a mais do que a política de tratar todos os pacientes, apresentando, assim, baixa utilidade clínica.

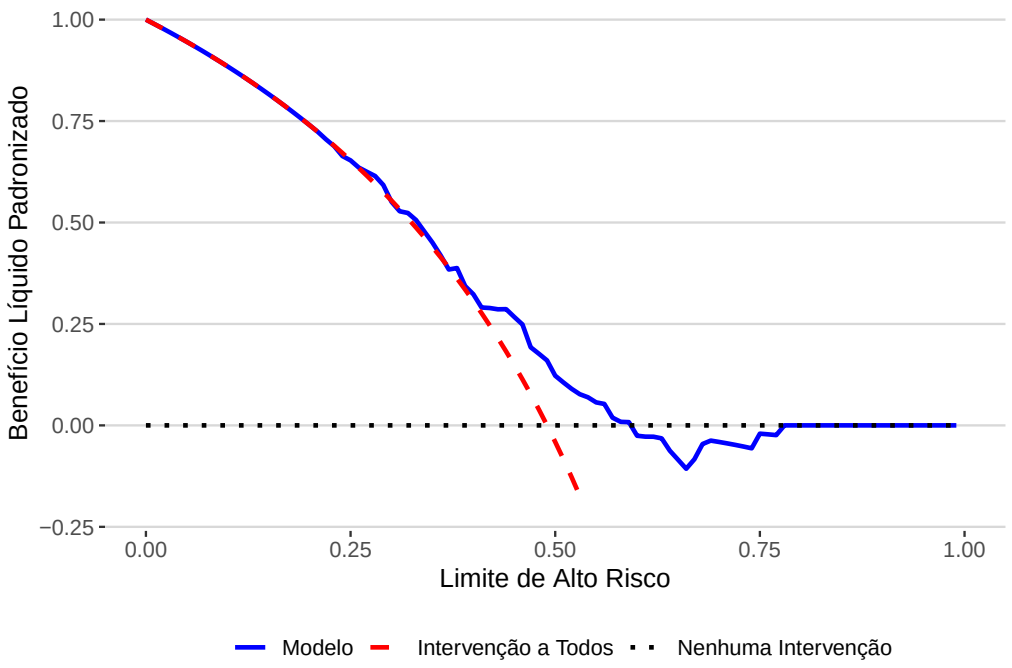

Figura 2.3: Exemplo de uma curva de decisão com dados simulados cujo modelo preditivo apresenta um percentual de área abaixo da curva ROC de 60,6 por cento. A curva azul mostra o benefício líquido ao se utilizar o modelo prognóstico, enquanto a curva vermelha representa a política de intervenção a todos.

A interpretação correta da curva de decisão é crucial para identificar a utilidade de um modelo prognóstico. Para isso, é necessário ressaltar que o benefício é a capacidade do modelo identificar corretamente pacientes com ou sem o evento. Um modelo preditivo com área abaixo da curva ROC de 50% apresenta curva de decisão idêntica à política "tratar

todos". Além disso, ao comparar curvas de decisão de diferentes modelos, o modelo de melhor ajuste terá sua curva acima das demais em toda a escala de possíveis pontos de corte, ou seja, apresentará maior benefício para qualquer ponto de corte escolhido.

A curva de decisão tem o intuito de avaliar e comparar os benefícios de adotar um modelo como suporte na decisão clínica, entretanto, a escolha do ponto de corte pode variar dependendo da preocupação médica e propensão à intervenção. Um médico pode estar mais preocupado com a doença e avaliar o risco de perder um diagnóstico de uma doença em maior grau e priorizar a busca por qualquer paciente em um estágio curável: pode ser adotado um ponto de corte menor para o diagnóstico. Um médico preocupado com o tratamento (intervenção) ser invasivo ou com o risco do paciente ao procedimento pode optar por tratar apenas aqueles com um risco particularmente alto, ou seja, adotando um ponto de corte maior (Vickers et al., 2019).

As técnicas de avaliação de modelos prognósticos apresentadas serão utilizadas para validar o escore $ABC_2 - SPH$ construído para pacientes com COVID-19 do Brasil (Validação Temporal) e da Espanha (Validação Externa). Os métodos foram aplicados conforme diretrizes estabelecidas pelo *PROBAST* (Wolff et al., 2019). Todas as análises foram realizadas utilizando o *software R* versão 4.0.2 e *Rstudio* versão 1.3.1093 e adotou-se 5% de nível de significância. Todos os intervalos de confiança apresentados foram construídos a partir de reamostragem *bootstrap* com 2000 replicações e correspondem ao nível de 95% de confiança.

3 Resultados e Discussão

Voltando ao caso em estudo, a base de dados utilizada foi separada em base de derivação (3978 pacientes - 79, 1%) e validação (1054 pacientes - 20, 9%) de acordo com a data de admissão do paciente, permitindo uma validação temporal. As duas bases foram imputadas independentemente utilizando MICE de forma que a base de derivação utilizou para a imputação, além das demais variáveis selecionadas, o desfecho, diferentemente da base de validação. Dessa forma, resultaram $m = 10$ bancos de dados para derivação e outros 10 para validação. A imputação múltipla adiciona uma certa complexidade para a análise como um todo, inclusive para a avaliação do modelo prognóstico desenvolvido. É necessário estabelecer como a validação do mesmo pode ser performada de maneira conveniente e que ofereça certa segurança aos resultados. É relevante destacar que é muito comum que não seja detalhado em artigos publicados como as estimativas dos bancos imputados foram combinadas (Marshall et al., 2009), o que provoca grande incerteza sobre qual a forma mais adequada de tratar tal problema. Contudo, algumas recomendações podem ser encontradas na literatura.

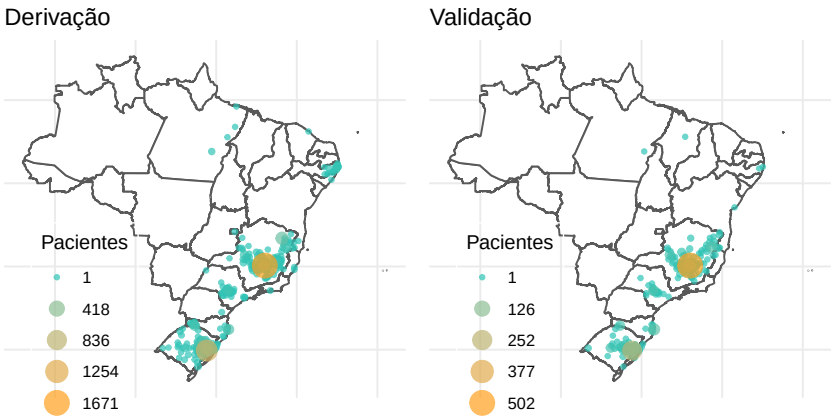

Figura 3.1: Município de residência dos pacientes utilizados para a construção (esquerda) e validação (direita) do escore $ABC_2 - SPH$.

3.1 Derivação do Escore $ABC_2 - SPH$

Utilizando as regras de Rubin (Rubin, 1988), o modelo prognóstico que explica a mortalidade em decorrência da COVID-19 a partir do escore desenvolvido foi ajustado em cada um dos m bancos imputados de derivação e os coeficientes resultantes foram combinados a partir da média simples. Estes coeficientes combinados são utilizados para a aplicação do escore e serão utilizados para validação.

3.2 Abordagens para Validação

Para aplicar as técnicas de avaliação do modelo em um cenário de múltiplos dados imputados é possível utilizar de três abordagens. A primeira consiste em avaliar o desempenho do modelo em cada um dos m bancos imputados de validação e, após isso, combinar as medidas de avaliação. Entretanto, há uma dificuldade e incerteza em qual a melhor forma de combinar as medidas de avaliação, especialmente em análises que requerem a avaliação gráfica como a curva de decisão. A segunda abordagem propõe combinar as m predições dos bancos de dados imputados de validação - para cada paciente - pela média das mesmas, sob o requisito das imputações das bases de derivação e validação terem sido realizadas de forma independente e o desfecho utilizado apenas no primeiro caso (Wood et al., 2015). Dessa forma, seria possível obter de maneira simples medidas únicas de avaliação. Outra possibilidade é avaliar o desempenho do modelo utilizando apenas os casos completos da base de validação, ou seja, somente os pacientes selecionados para validação que possuem na base de dados não-imputada todas as informações necessárias para o cálculo do escore. Esta visão é importante, também, para identificar um possível viés que a imputação possa estar inserindo no estudo. Todas as três abordagens serão utilizadas para a avaliação do modelo prognóstico em estudo e os seus resultados serão comparados.

É importante mencionar que as bases de derivação do modelo e validação possuem percentual de óbitos semelhantes: $20,26\%$ e $19,73\%$, respectivamente. Enquanto que os casos completos da base de validação compõem 779 pacientes, com incidência de óbito de $18,99\%$.

3.3 Validação Temporal

O escore prognóstico desenvolvido para estratificar o risco de pacientes com COVID-19 obteve ótima discriminação tanto para os casos completos quanto ao combinar as predições, como pode ser visto na Figura 3.2. Resultaram áreas abaixo da curva ROC de $85,2\%$ e $85,9\%$, respectivamente.

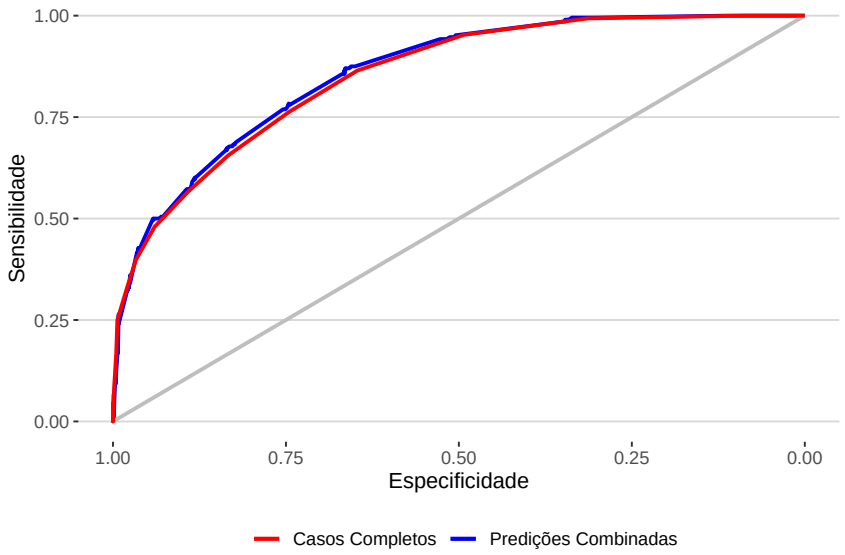

Figura 3.2: Curvas ROC obtidas ao utilizar apenas os casos completos de validação (vermelho) e combinar as predições dos 10 bancos de validação imputados (azul).

A Figura 3.3 mostra o escore de Brier obtido de $0,108$ ao combinar as predições e $0,107$ para os casos completos, indicando um bom desempenho geral.

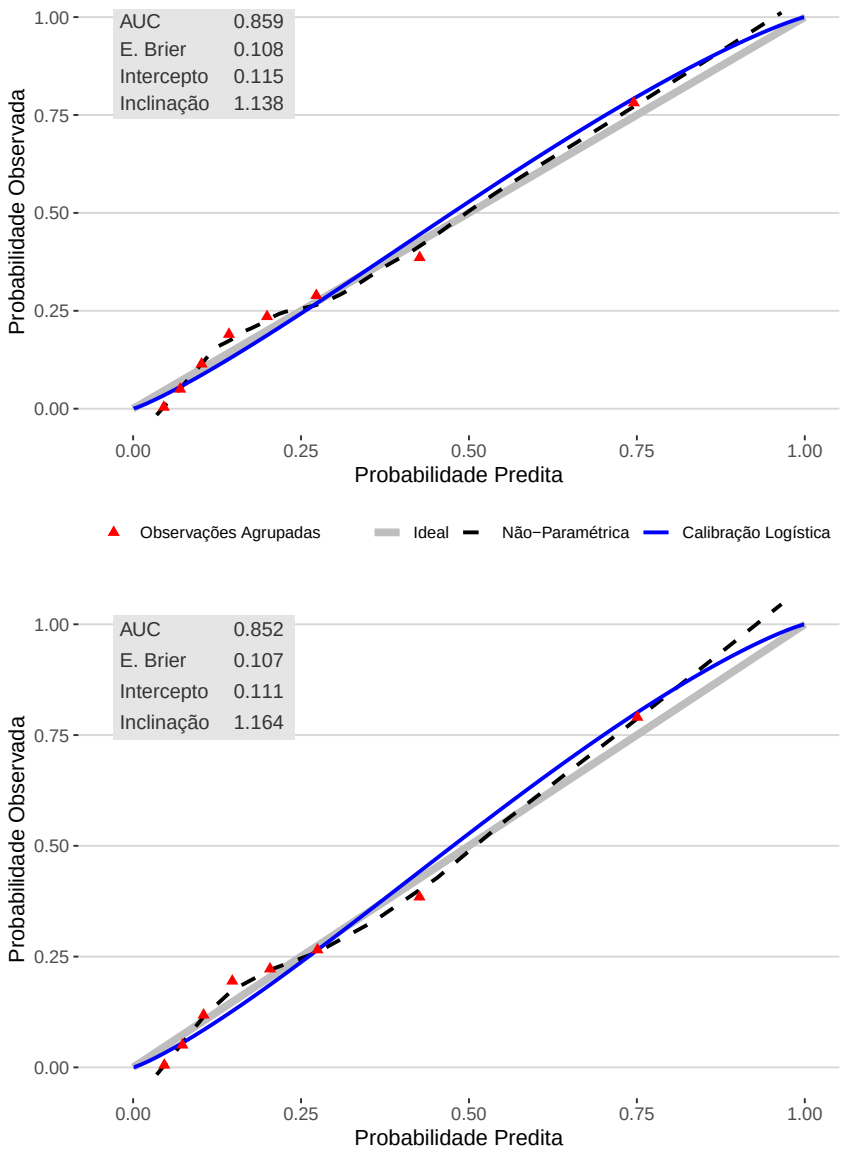

Figura 3.3: Curvas de calibração logística (em azul) e não-paramétrica (preta) obtidas ao combinar as predições dos bancos imputados de validação (acima) e ao utilizar apenas os casos completos (abaixo). No canto superior esquerdo dos gráficos estão algumas métricas de avaliação do modelo. Os pontos em vermelho mostram 8 grupos de pacientes agrupados pelos quantís de mortalidade.

Quanto à calibração, a Figura 3.3 mostra que tanto a curva logística quanto a não-paramétrica de calibração estão muito próximas à reta ideal para ambas as abordagens apresentadas nos gráficos. Ao combinar as predições, as duas retas indicam uma leve subestimação da probabilidade de pacientes com probabilidade observada entre 50% e 90% (calibração fraca). Enquanto que os grupos de pacientes e a curva não-paramétrica ainda mostram leve superestimação dos pacientes com os menores percentuais de óbito observados, seguido por uma subestimação que engloba do 3º ao 5º grupo de quantís (calibração moderada). Para os casos completos, os resultados são muito semelhantes. Outro indício de boa calibração são o intercepto e inclinação de Cox com valores observados de $0,115$ e

$1,138$ ao combinar as predições e $0,111$ e $1,164$ para os casos completos, respectivamente, e com o teste de hipóteses $H_0 : Int = 0, Inc = 1$ retornando *p-valor* $0,17$ para os casos completos e $0,185$ para as predições combinadas.

No geral, a calibração observada com os dados de validação imputados e completos é muito boa, apresentando apenas pequenos desvios da reta ideal. Além disso, é possível observar grande concentração de pacientes com até 25% de mortalidade predita ou observada.

As curvas de decisão na Figura 3.4 mostram que o escore desenvolvido apresenta ótima utilidade clínica, oferencendo grandes benefícios com relação às políticas de intervenção a todos ou a nenhum paciente. Foi observada leve variação entre as curvas dos casos completos e de validação imputados.

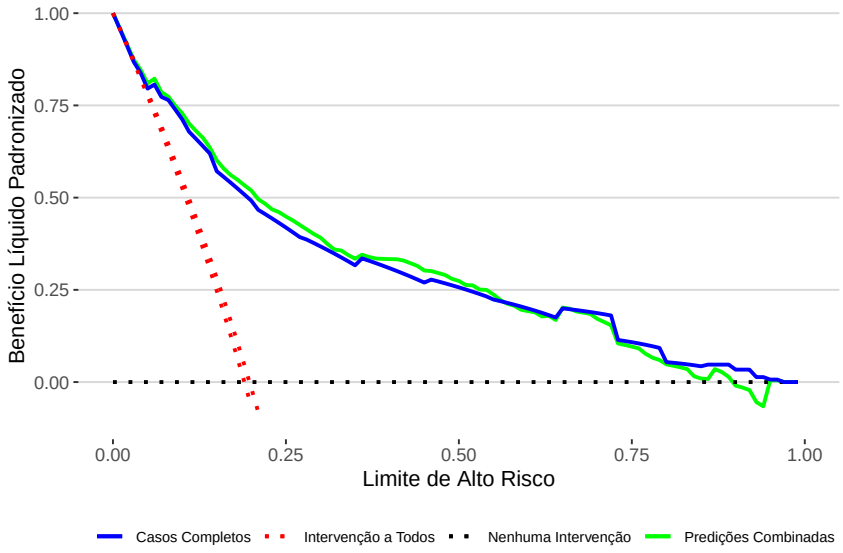

Figura 3.4: Curvas de decisão para o escore prognóstico desenvolvido para pacientes internados com COVID-19. A curva azul mostra o benefício líquido padronizado do modelo para diferentes pontos de corte para os casos completos de validação e a curva verde ao se combinar as predições dos 10 bancos de validação imputados.

A Tabela 3.1 mostra todas as métricas calculadas nas 3 abordagens utilizadas. Ao aplicar as técnicas de avaliação em todos os bancos imputados são obtidos 10 medidas por métrica e elas são combinadas por média simples, mas os valores mínimos e máximos também são exibidos. Em geral, os valores são muito similares entre as diferentes abordagens e todas as métricas mostram um ótimo desempenho tanto em termos de desempenho global quanto de calibração e discriminação, agregando confiança à validação e ao escore desenvolvido.

Com relação à discriminação do escore em sua aplicação, deve-se considerar os grupos de risco construídos para auxiliar a tomada de decisão e alocação de recursos clínicos. A Tabela 3.2 apresenta as probabilidades preditas pelo modelo utilizadas para estratificar o risco dos pacientes em 4 grupos e os percentuais de mortalidade observados nas bases de validação imputada e de casos completos. Para a base imputada, cada paciente obtém 10 escores (um correspondente a cada base imputada) e, destes, foi considerado como escore definitivo o mais frequente. Em caso de empate, selecionou-se o maior escore. É possível observar que os percentuais observados na base são coerentes com os cortes do risco predito utilizados na estratificação. Com isso, é possível obter grupos de risco de

Tabela 3.1: Métricas de Avaliação por Abordagem

	Bancos Imputados	Predições Combinadas	Casos Completos
AUC	85,69% (85,40%-85,92%)	85,86% (83,26%-88,47%)	85,2% (82,04%-88,36%)
Escore de Brier	0,109 (0,108-0,109)	0,108	0,107
Intercepto	0,102 (0,086-0,132)	0,115	0,111
Inclinação	1,127 (1,115-1,148)	1,138	1,164
P-valor (Int=0, Inc=1)	23,59% (15,61%-29,02%)	18,48%	16,99%

Nas métricas de avaliação por banco de validação imputado são exibidos: valor médio (mínimo - máximo). Para as predições combinadas e casos completos são exibidos os valores observados e intervalo de 95% de confiança para a AUC.

Tabela 3.2: Grupos de Risco do Escore $ABC_2 - SPH$ por Base de Validação

		Bases Imputadas		Casos Completos	
Grupo de Risco	Prob. Predita	Pacientes	Óbitos	Pacientes	Óbitos
Baixo (0-1)	0%-5,9%	290	1 (0.3%)	199	1 (0.5%)
Intermediário (2-4)	6%-14,9%	394	47 (11.9%)	306	34 (11.1%)
Alto (5-8)	15%-49,9%	252	73 (29%)	194	54 (27.8%)
Muito Alto (9-20)	50%-100%	118	87 (73.7%)	80	59 (73.8%)
Geral	-	1054	208 (19.7%)	779	148 (19%)

Para as bases imputadas foi obtido o escore mais frequente (moda) do paciente.

pacientes admitidos com COVID-19 utilizando o escore $ABC_2 - SPH$ que possam auxiliar de maneira efetiva e confiável a alocação de recursos.

3.4 Validação Externa

Com a validação temporal realizada é possível certificar com certa segurança que o escore prognóstico de estratificação de risco desenvolvido para pacientes com COVID-19 apresenta performance adequada e possui grandes benefícios para ser aplicado na população em estudo (pacientes com COVID-19 admitidos em hospitais públicos e privados no Brasil). A abrangência do estudo pode ser visualizado na Figura 3.1 que mostra o município de residência dos pacientes que participaram da construção do escore separados por derivação e validação. Entretanto, para que o escore desenvolvido possa ser aplicado em outras populações é necessário avaliar a performance do mesmo em outras amostras de pacientes que compreendem outras populações similares.

474 pacientes admitidos com COVID-19 no Hospital Universitário Vall d'Hebron, em Barcelona, Espanha, entre março e maio de 2020 foram selecionados para a aplicação e validação do escore $ABC_2 - SPH$. Todos estes pacientes contém as informações necessárias para o cálculo do escore. A amostra contém 82 óbitos (incidência de $17,3\%$) e os critérios de inclusão e exclusão de pacientes foram os mesmos adotados para a construção do escore.

A Figura 3.5 mostra que foi obtido um escore de Brier de $0,093$ indicando bom desempenho geral e uma área abaixo da curva ROC de $89,9\%$ (IC: $86,4\%$ - $93,4\%$), mostrando que o escore foi muito efetivo ao discriminar os pacientes com relação à mortalidade. Entretanto, as curvas de calibração mostram que o escore superestima a probabilidade de óbito dos pacientes com baixo percentual observado e, principalmente, subestima a probabilidade

daqueles com percentual observado maior que 20%. Isto resultou em valores do intercepto e inclinação de $0,73$ e $1,52$, com p-valor $0,001$ que rejeita a hipótese $H_0 : Int = 0, Inc = 1$.

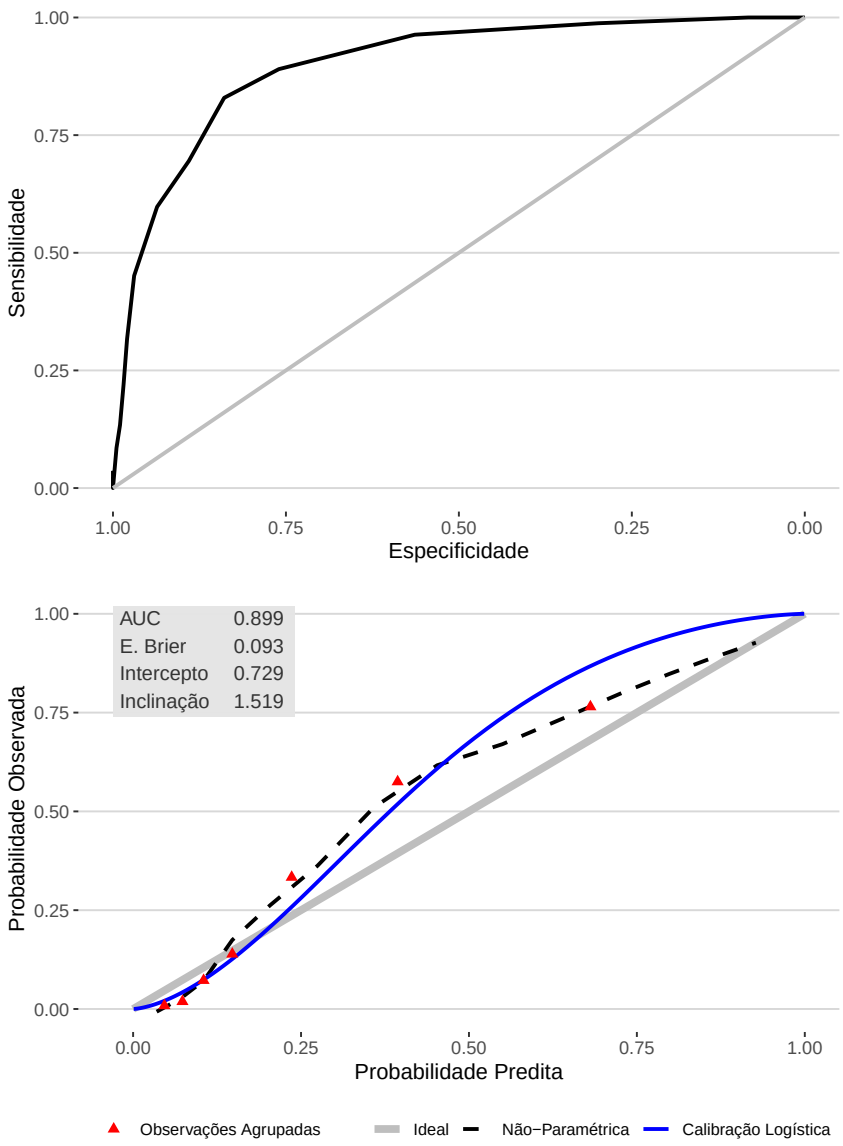

Figura 3.5: Medidas de validação externa para o escore $ABC_2 - SPH$. Acima: Curva ROC. Abaixo: Curvas de Calibração.

A Tabela 3.3 mostra como os percentuais de óbitos observados para os grupos de risco "Baixo" e "Intermediário" são consideravelmente inferiores aos percentuais observados nas bases de validação imputadas, enquanto que, fazendo a mesma comparação, os grupos "Alto" e "Muito Alto" apresentam percentualis de óbitos superiores na base de validação externa. Essas diferenças observadas nos grupos de risco refletem o que também é observado no gráfico de calibração. Apesar destas diferenças, o escore conseguiu discriminar muito bem os pacientes com menor risco e maior risco de óbito em decorrência da doença causada pelo coronavírus.

A curva de decisão do escore $ABC_2 - SPH$ aplicado nos pacientes espanhóis, Figura 3.6, mostra que há grandes benefícios ao se utilizar o escore construído em comparação com as políticas de intervenção a todos e a nenhum paciente.

Tabela 3.3: Grupos de Risco do Escore $ABC_2 - SPH$ nas Bases de Validação Imputadas e Validação Externa

		Bases Imputadas		Validação Externa	
Grupo de Risco	Prob. Predita	Pacientes	Óbitos	Pacientes	Óbitos
Baixo (0-1)	0%-5,9%	290	1 (0.3%)	118	1 (0.008%)
Intermediário (2-4)	6%-14,9%	394	47 (11.9%)	225	13 (5.8%)
Alto (5-8)	15%-49,9%	252	73 (29%)	97	42 (43.3%)
Muito Alto (9-20)	50%-100%	118	87 (73.7%)	34	26 (76.5%)
Geral	-	1054	208 (19.7%)	474	82 (17.3%)

Para as bases imputadas de derivação foi obtido o escore mais frequente (moda) do paciente.

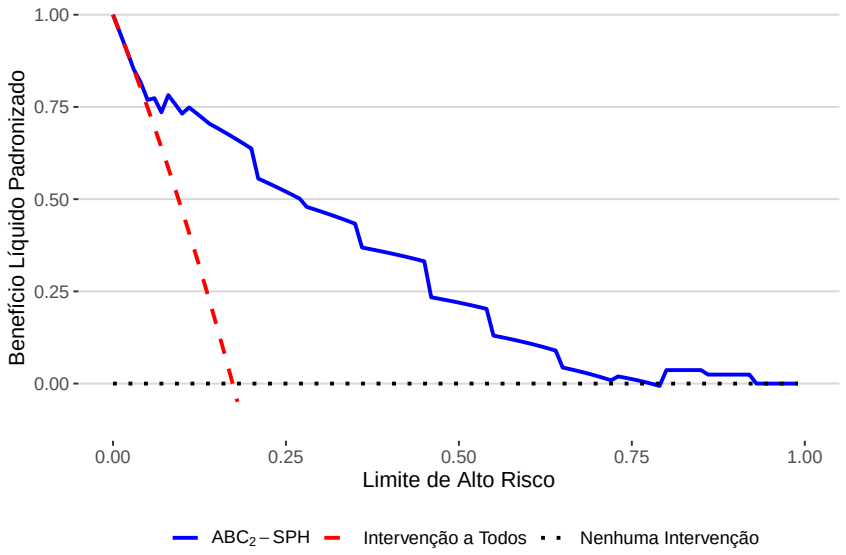

Figura 3.6: Curva de decisão da validação externa para o escore $ABC_2 - SPH$. As linhas azul, vermelha e preta representam, respectivamente, o benefício líquido padronizado ao se utilizar o escore, ao aplicar tratamento a todos e a nenhum paciente.

Dessa forma, o escore $ABC_2 - SPH$ obteve métricas de avaliação que garantem ótima discriminação dos pacientes com COVID-19 com relação à mortalidade e grande utilidade clínica, o que na prática pode ser o suficiente para que o mesmo seja utilizado a partir dos grupos de risco. Contudo, para um melhor ajuste, é possível fazer uma recalibração do escore, corrigindo o intercepto e inclinação do modelo (recalibração logística) para que se adequem aos pacientes de Barcelona (Sim et al., 2016) e, assim, os percentuais de óbitos preditos pelo modelo sejam reajustados.

É importante destacar que questões como o tamanho amostral, assim como o fato de terem sido considerados pacientes de apenas um centro médico de Barcelona, podem ter influenciado nas métricas de validação externa. O *PROBAST* recomenda que existam ao menos 100 eventos para validar um modelo prognóstico. Outro ponto relevante a se destacar está nas diferentes políticas de internação dos pacientes no Brasil e Espanha durante a pandemia do novo coronavírus: o país espanhol adotou a política de internar os pacientes no estágio inicial da doença o que refletiu em uma mediana de 21 dias de internação dos pacientes pertencentes à amostra utilizada na validação externa, enquanto que no Bra-

sil a base de validação obteve mediana de 7 dias de internação até o desfecho. Como o escore $ABC_2 - SPH$ utiliza variáveis correspondentes à admissão do paciente no centro médico, os pacientes espanhóis podem ter sido internados com um padrão de variáveis distinto o suficiente para requerer uma recalibração para que o escore seja aplicável a essa população.

4 Conclusões

Após a avaliação do escore $ABC_2 - SPH$ para pacientes com COVID-19 com relação à discriminação, calibração e utilidade clínica tanto para pacientes do Brasil quanto para pacientes de Barcelona, foi possível validar com certa segurança a utilização do escore desenvolvido para estratificar o risco de óbito de pacientes infectados com o novo coronavírus no Brasil em hospitais públicos e privados. A validação temporal realizada demonstrou ótima discriminação feita pelo escore com relação ao risco de óbito intra-hospitalar dos pacientes, além de boa calibração e grandes benefícios ao se utilizar esse método com relação aos métodos de tratamento a todos ou a nenhum paciente.

Com relação à utilização do escore em outras populações, a validação externa realizada com pacientes de um centro médico de Barcelona, Espanha, demonstrou excelente discriminação no escore $ABC_2 - SPH$ ao estratificar o risco de óbitos dos pacientes com COVID-19, além de indicar grandes benefícios ao utilizá-lo no lugar das políticas padrão de tratamento. Entretanto, a análise da calibração do escore para estes pacientes mostrou que o modelo prognóstico desenvolvido pode subestimar ou superestimar as probabilidades de óbito dos pacientes, sendo necessária uma recalibração do escore para melhor resultados e para se adequar a esses pacientes. De toda forma, o escore $ABC_2 - SPH$ pode ser utilizado para discriminar os pacientes espanhóis a partir dos grupos de risco de óbito (Baixo, Intermediário, Alto e Muito Alto) como mostrado na análise.

Em estudos futuros, a recalibração do escore $ABC_2 - SPH$ pode ser feita para se obter um novo escore prognóstico que estime de forma mais precisa a probabilidade de óbito dos pacientes espanhóis. Outra vertente de estudo seria a aplicação e validação dos escore $ABC_2 - SPH$ em outras populações, preferencialmente com amostras maiores e que abranjam diversos centros médicos. E assim, o escore $ABC_2 - SPH$ possa ser validado ou mesmo recalibrado para que possa trazer benefícios à outras populações, auxiliando na decisão clínica e na alocação de recursos nos centros médicos a fim de garantir o tratamento necessário àqueles que o necessitam.

Referências Bibliográficas

Altman, D. G., Vergouwe, Y., Royston, P., and Moons, K. G. (2009). Prognosis and prognostic research: validating a prognostic model. *Bmj*, 338:b605.

Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., and Conde, J. G. (2009). Research electronic data capture (redcap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of biomedical informatics*, 42(2):377–381.

Huang, Y., Li, W., Macheret, F., Gabriel, R. A., and Ohno-Machado, L. (2020). A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association*, 27(4):621–633.

Marcolino, M. S., Pires, M. C., Ramos, L. E. F., Silva, R. T., Oliveira, L. M., Carvalho, R. L., Mourato, R. L., Sánchez-Montalvá, A., Raventos, B., Anschau, F., et al. (2021). Abc2-sph risk score for in-hospital mortality in covid-19 patients: development, external validation and comparison with other available scores. *medRxiv*.

Marshall, A., Altman, D. G., and Holder, R. L. (2009). Combining estimates of interest in prognostic modelling.

Rubin, D. B. (1988). An overview of multiple imputation. In *Proceedings of the survey research methods section of the American statistical association*, pages 79–84. Citeseer.

Sim, J., Teece, L., Dennis, M. S., Roffe, C., and Team, S. S. (2016). Validation and recalibration of two multivariable prognostic models for survival and independence in acute stroke. *PLoS One*, 11(5):e0153527.

Steyerberg, E. W. et al. (2019). *Clinical prediction models*. Springer.

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., and Steyerberg, E. W. (2019). Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1):1–7.

Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M. J., and Steyerberg, E. W. (2016). A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of clinical epidemiology*, 74:167–176.

Vickers, A. J., van Calster, B., and Steyerberg, E. W. (2019). A simple, step-by-step guide to interpreting decision curve analysis. *Diagnostic and prognostic research*, 3(1):1–8.

Wolff, R. F., Moons, K. G., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., and Mallett, S. (2019). Probast: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of internal medicine*, 170(1):51–58.

Wood, A. M., Royston, P., and White, I. R. (2015). The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. *Biometrical Journal*, 57(4):614–632.