Universidade do Estado do Rio de Janeiro

Centro Biomédico

Instituto de Medicina Social

Marisa da Silva Santos

Validação externa de modelos de predição de pneumonia pós cirurgia cardíaca

Rio de Janeiro

2010

Marisa da Silva Santos

Validação externa de modelos de predição de pneumonia pós cirurgia cardíaca

Tese apresentada, como requisito parcial para obtenção do título de Doutor, ao Programa de Pós-Graduação em Saúde Coletiva do Instituto de Medicina Social, da Universidade do Estado do Rio de Janeiro. Área de concentração: Epidemiologia.

Orientador: Prof. Dr. José Ueleres Braga

Rio de Janeiro

2010

CATALOGAÇÃO NA FONTE
UERJ/REDE SIRIUS/CBC

S237 Santos, Marisa da Silva.
Validação externa de modelos de predição de pneumonia pós cirurgia cardíaca / Marisa da Silva Santos. – 2010.
90f.

Orientador: José Uereles Braga.
Tese (Doutorado) – Universidade do Estado do Rio de Janeiro, Instituto de Medicina Social.

1. Predição (Lógica) – Teses. 2. Pneumonia – Teses. 3. Coração – Cirurgia – Teses. 4. Prognóstico – Métodos estatísticos – Teses. I. Braga, José Uereles. II. Universidade do Estado do Rio de Janeiro. Instituto de Medicina Social. III. Título.

CDU 616.24-002

_____ _____

Assinatura Data

Marisa da Silva Santos

Validação externa de modelos de predição de pneumonia pós cirurgia cardíaca

Tese apresentada, como requisito parcial para obtenção do título de Doutor, ao Programa de Pós-Graduação em Epidemiologia do Instituto de Medicina Social, da Universidade do Estado do Rio de Janeiro. Área de concentração: Epidemiologia.

Aprovada em: 6 de abril de 2010.

Banca Examinadora:

_____

Prof. Dr. José Ueleres Braga (Orientador)

Instituto de Medicina Social - UERJ

_____

Prof. Dr. Paulo Feijó Barroso

Universidade Federal do Rio de Janeiro- UFRJ

_____

Prof. Dr. Antonio José Leal Costa

Universidade Federal do Rio de Janeiro- UFRJ

_____

Prof.ª Dra. Gulnar Azevedo e Silva Mendonça

Instituto de Medicina Social – UERJ

_____

Prof. Dr. Guilherme Loureiro Werneck

Instituto de Medicina Social – UERJ

Rio de Janeiro

2010

DEDICATÓRIA

À Fernanda, meu amor por você é maior que o infinito.

AGRADECIMENTOS

Ao meu marido pela sua ajuda e apoio em todos os momentos.

Ao Prof. Jelle Goeman e ao prof. Ewout Steyerberg pela sua ajuda na elaboração da rotina em R.

Aos amigos do Instituto Nacional de Cardiologia, em especial ao prof. Bernardo Tura e as equipes da UTCIC e da CCIH.

Aos novos amigos da SESDEC onde tenho aprendido muito e conseguido entender o Sistema de Saúde Público.

E por último ao meu orientador Prof. Ueleres pela sua competência e disponibilidade.

RESUMO

SANTOS, Marisa da Silva. *Validação externa de modelos de predição de pneumonia pós cirurgia cardíaca.* 2010. 90f. Tese (Doutorado em Saúde Coletiva) – Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2010.

Este trabalho versa sobre a validação externa de um modelo para predição de pneumonias em pacientes submetidos a cirurgias cardíacas. Também apresenta uma revisão dos métodos e técnicas para análise crítica e avaliação de desempenho dos modelos preditivos em medicina e discorre sobre aplicações do nomograma. Método: Dados de uma série de 527 pacientes, consecutivamente submetidos a cirurgias cardíacas entre Junho de 2000 e Agosto de 2002, foram utilizados para desenvolver os modelos de prognósticos. Foram realizadas análise de regressão logística múltipla e árvore de classificação e regressão (CART) para identificar fatores preditivos para a ocorrência de pneumonia. Diversos fatores de risco simples e convencionais pré-operatórios foram avaliados. Os modelos foram validados internamente com um método de *bootstrap* Um nomograma foi desenvolvido para melhorar a aplicabilidade clínica. O desempenho do nomograma foi avaliado por meio de medidas de calibração, discriminação e indicadores globais. Em uma segunda etapa estudo em um hospital público foi realizado com 333 pacientes adultos submetidos a cirurgias cardíacas entre Outubro de 2006 e Maio de 2007. Modelos construídos previamente por meio de regressão logística (LRM) e árvore de classificação e regressão (CART) foram validados com dados externos. Resultados: Um modelo de nomograma simples foi desenvolvido e validado internamente, mostrando discriminação moderada e boa calibração (AUC 0,79; escore Brier 0,064, ângulo de discriminação 0,13; *Hosmer-Lemeshow* p = 0,27). Pneumonia ocorreu em 7,6% dos pacientes da amostra de validação externa. LRM apresentou melhor desempenho com baixa discriminação (R^2 7,1%, Brier=0,06, AUC=0,694) e com calibração adequada (*Hosmer-Lemeshow* p=0,08). Conclusões: As probabilidades preditas mostraram concordância global com a freqüência observada de pneumonia após cirurgia cardíaca. O nomograma forneceu uma predição satisfatória da probabilidade de pneumonia. Sua aplicabilidade para o uso clínico pode facilitar a informação do paciente e do cirurgião antes da cirurgia cardíaca. Foi validado externamente um modelo capaz de identificar pacientes de alto risco para pneumonia submetidos à cirurgias cardíacas. CART apresentou um bom desempenho na derivação e maiores perdas do que LRM, quanto à discriminação e calibração, na amostra de validação.

Palavras-chave: Pneumonia. Métodos epidemiológicos. Modelos de predição. Cirurgia cardíaca. Estudos de validação.

ABSTRACT

This study concerns the external validation of a prediction model for pneumonias after cardiac surgery. It also presents a review of methods and techniques for critical appraisal and performance assessment of clinical predictive models and nomogram applications. Methods: A consecutive series of 527 patients who underwent cardiac surgeries between June 2000 and August 2002 was used to develop a prognostic model.. Multiple logistic regression analysis was performed to predict the occurrence of pneumonia. Diverse simple and conventional preoperative risk factors were evaluated. The model was internal validated with bootstrap. A nomogram was developed to enhance clinical applicability. The performance was evaluated by calibration, discrimination and global measures. Prospective study was done to validate models predicting pneumonia after cardiac surgery with 333 adult patients who underwent cardiac surgery from October 2006 to May 2007. Previously constructed logistic regression (LRM) and classification and regression tree (CART) models were validated with external data. Results: a simple nomogram model was developed and showing low discrimination and good calibration (AUC 0.79, Brier score 0.064, discrimination slope 0.13, Hosmer-Lemeshow p=0.27). Pneumonia occurred in 7.5% of patients in the external validation set. LRM performed better with moderate discrimination (R2 7.1%, Brier=0.06, AROC=0.694) and calibration (Hosmer-Lemeshow P=0.08). Conclusions: Overall agreement between the predicted probabilities and observed frequencies was good in the development and the internal validation set. The nomogram predicts the probability of pneumonia for individual patients and may help in informing patients and surgeons before undergoing cardiac surgery. We validated a model that can identify which patients undergoing cardiac surgery are at high risk for pneumonia. CART performs well in derivation, and looses more discrimination and calibration than LRM in the validation set.

Keywords: Pneumonia. Validation studies. Epidemiologic methods. Prediction models. Cardiac surgical procedures.

LISTAS DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

AID	*Automatic Interation Detection*- Detecção automática de interações
ASA	*American Society of Anestesiology* - Sociedade Americana de Anestesiologia
AUC	*Área Under Curve* [ROC]- Área sob a curva ROC
AVC	Acidente Vascular Encefálico
CAHTA	*Catalan Agency for Health Technology Assessment*- Agencia Catalã de Avaliação de Tecnologias em Saúde
CART	*Classification and Regression Tree* - Árvore de classificação e regressão
CDC	*Centers for Disease Control*- Centros para Controle de Doenças
CEC	Circulação extra- corpórea
DPOC	Doença Pulmonar Obstrutiva Crônica
ECG	Eletrocardiograma
EUROSCORE	*European System for Cardiac Operative Risk Evaluation*- Sistema Europeu para Avaliação de Risco Cirúrgico Cardíaco
IAM	Infarto Agudo do Miocárdio
IFO	Infecção de Ferida Operatória
IMC	Índice de Massa Corporal
Kg	Kilograma
LRM	Regressão Logística Múltipla
MACE	Major Cardiovascular Events – Eventos Cardiovasculares Maiores
NNE	*Northern New England*- Nova Inglaterra Norte
NNIS	*National Nosocomial Infection Surveillance* – Vigilância Nacional de Infecções Hospitalares
NYHA	*New York Heart Association*- Associação do Coração de Nova Iorque
ROC	*Receiver-operating characteristic curve*
RPC	Regra de Predição Clínica
RVM	Revascularizaçao do Miocárdico
UTI	Unidade de Terapia Intensiva

SUMÁRIO

INTRODUÇÃO

Apresentação

O trabalho se divide nas seguintes partes:

A seção *Introdução* discute o impacto da cirurgia cardíaca no Brasil, complicações mais freqüentes, pneumonia hospitalar e sua prevenção. Também revisa as normas para construção de modelos preditivos e revisa os modelos de maior importância em infecção e em cirurgia cardíaca. Apresenta os objetivos gerais da tese contemplados em três artigos.

Na seção *Metodologia* são apresentados o fluxograma da pesquisa, e os detalhes da definição e mensuração das variáveis preditoras e diagnóstico do desfecho.

Na seção *Artigo I* é apresentada uma revisão sobre as diversas técnicas e métodos para análise crítica e avaliação de desempenho dos modelos preditivos em medicina.

Na seção *Artigo II* são apresentados os detalhes da construção de um nomograma para o modelo logístico, sua validação interna e a mensuração do seu desempenho por diversas técnicas.

Na seção *Artigo III* os modelos de árvore de classificação e logístico são submetidos ao processo de validação externa e o seu desempenho é comparado entre si e aos valores ideais por métodos gráficos e empíricos.

Em *Resultados Complementares* são apresentados gráficos comparativos entre a derivação e a validação externa para os modelos de árvore e regressão logística não utilizada nos artigos.

Na seção *Considerações Finais* são resumidos os achados dos três artigos, resultados obtidos e sugestões para futura continuação da pesquisa.

Na seção *Referências* são listados os artigos citados no corpo do texto da tese.

Em *Anexo I é* demonstrada uma rotina completa em R para derivação modelo logístico e árvore, criação de nomograma, validação interna e externa construído para a tese.

Na seção *Anexo II* é apresentada a rotina para Teste *Hosmer-Lemeshow* desenvolvida e disponibilizada pelo prof Steyerberg da Universidade de Roterdan.

Em *Anexo III* são disponibilizadas as rotinas para Teste de *Goeman le Cessie* desenvolvidas pelo prof. Jelle Goeman da Universidade de Leiden.

Cirurgia Cardíaca: complicações e predisposição à pneumonia

A cirurgia cardíaca vem crescendo em volume no Brasil. Segundo dados do Sistema de Informações Hospitalares [1] no ano de 2008 foram realizados 64.519 cirurgias cardíacas, totalizando cerca de 570 milhões em gastos, com concentração de procedimentos na Região Sudeste (quadro 1):

Cirurgias Cardíacas do SUS - por local de internação - Brasil
Valor total por Ano competência segundo Região

Região	*2008*		*2009*	
	n procedimentos	*Valor Total*	*n procedimentos*	*Valor Total*
Região Norte	1936	17,585,635.2	2181	21,133,155.4
Região Nordeste	11169	89,541,849.9	12280	109,111,674.1
Região Sudeste	31485	275,640,073.7	34541	329,793,425.7
Região Sul	14811	141,559,960.7	15081	155,820,910.7
Região Centro-Oeste	5129	45,641,000.1	5502	55,763,710.0
Total	64530	569,968,519.8	69585	671,622,876.2

Quadro 1 - Valor total cobrado pelas unidades hospitalares – SIH-DATASUS para cirurgias cardíacas no Brasil em 2008 e 2009

A cirurgia cardíaca apresenta ainda elevado percentual de complicações com uma letalidade esperada de até 11% e complicações de até 50%, em especial infecções, sangramentos, complicações neurológicas, insuficiência renal e isquemia mesentérica [2]. As cirurgias cardíacas são habitualmente realizadas com circulação extracorpórea (CEC). A CEC é causa de imunodeficiência humoral, de redução da fagocitose e da síndrome pós-perfusão, entidade esta caracterizada por inflamação e disfunção multiorgânica, podendo levar a danos pulmonares [3]. As cirurgias cardíacas são de duração prolongada normalmente três a quatro horas, permanecendo o paciente em CEC durante 30 a 120 minutos. Durante o período de CEC o pulmão não é ventilado, acarretando atelectasias e acúmulo de secreções.

O paciente é mantido em posição supina a zero grau durante o procedimento que pode durar mais de 5 horas, o que é considerado fator de risco importante para pneumonia [4]. Outro fator que pode agravar a evolução pós-operatória dos pacientes é a hipotermia que predispõe ao sangramento. A dor pós-operatória é freqüente e o uso de dois a três drenos é rotineiro, aumentando o risco da ocorrência de hipoventilação pulmonar. Em um procedimento sem intercorrências o paciente permanece em ventilação mecânica por pouco

tempo até que tenha despertado adequadamente, mas, em casos complicados, pode necessitar de ventilação mecânica por vários dias.

As infecções são as complicações mais freqüentes da cirurgia cardíaca, podendo chegar a 20% dos casos [5], sendo mais freqüentes as de ferida cirúrgica, a pneumonia e a sepse relacionada a cateteres vasculares. A pneumonia é a infecção hospitalar que mais leva ao óbito, com um risco relativo estimado de 14 [6], e também está associados a expressivos gastos, estimados de US$ 1450.00 a US$ 14 mil por paciente [7, 8], causados principalmente, pelo aumento no tempo de ventilação mecânica e de internação em UTI.

As medidas de prevenção de pneumonia nesta população ainda são pouco estudadas. A criação de modelos preditivos pode auxiliar no ajuste de risco, escolha de população para intervenções de maior custo e para compilar indiretamente o conhecimento sobre a fisiopatogenia da pneumonia nesta população. Os modelos de predição a serem estudados nesta tese foram originados de um primeiro estudo publicado em 2007[9] e testado por validação cruzada, com uma acurácia de 90%.

Modelos de predição clínica

A idéia de prever a possibilidade de ocorrência de um agravo faz parte da prática médica, habitualmente realizada de forma intuitiva a partir de conhecimento teórico prévio, exames físicos e dados laboratoriais mais simples.

As regras de predição clínica (RPC) têm objetivos variados como prognóstico, medir a probabilidade de eventos adversos e desfechos, classificar a gravidade da doença e elaborar estratificação de risco, de maneira a permitir comparações entre períodos ou instituições diferentes. Habitualmente, são desenvolvidas em situações complexas, incertas e de risco, ou quando existe uma possibilidade de reduzir custos sem comprometer a segurança do paciente [10]. Na prática, poucas RPC se arriscam a sugerir decisões, trabalhando mais a idéia de probabilidades por escores ou algoritmos.

Uma regra de predição clínica ideal [11-13] deve conter:

1. Desfecho claramente definido e clinicamente importante;
2. Avaliação cega do diagnóstico em relação aos preditores;
3. Variáveis de predição: devem ser claras e reprodutíveis, com uma lista de todas as variáveis avaliadas e não incluídas, mensuração cega das variáveis sem conhecer o desfecho;

4. População: todas as características que podem afetar o desfecho devem ser descritas;
5. Descrição do local do estudo;
6. Validação prospectiva em um grupo de pacientes diferente do grupo de derivação;
7. Análise de impacto - como o uso na prática afeta os desfechos clínicos;
8. Técnicas matemáticas bem descritas;
9. Descrição dos resultados;
10. Reprodutibilidade: confiabilidade interobservador das variáveis preditoras;
11. Sensibilidade: avaliação se a regra tem validade de face.

Modelos estudados em infecção hospitalar e cirurgia cardíaca

Foi realizada revisão da literatura utilizando as seguintes palavras-chave: "(predict* [tiab] OR predictive value of tests [mh] OR scor* [tiab] OR observ* [tiab] OR observer variation [mh]), C*ardiac Surgical Procedures", "Infection" and validation* or *validate,* na base Medline, e descritores semelhantes nas bases Lilacs, Scirus e Google acadêmico sendo selecionados estudos envolvendo infecções, óbitos e lesões pulmonares . Não houve exclusão de trabalhos com base na avaliação crítica. Foram excluídos modelos com finalidade única de identificar fatores de risco ou elaborar diagnósticos. Os trabalhos foram classificados pela autora utilizando os seguintes critérios adaptados de Gyatt et. al.[14]:

Nível I - localizado pelo menos um estudo de validação em nova população e um demonstrando impacto no seu uso no cuidado direto com o paciente.

Nível II - localizado pelo menos um estudo de validação em população diferente da original.

Nível III - localizado pelo menos um estudo de validação prospectiva.

Nível IV-localizado apenas estudos de derivação, validação cruzada, *bootstrap* ou amostras retrospectivas.

Poucos modelos foram construídos visando avaliar infecções hospitalares. Em cirurgias cardíacas a maioria dos modelos visa prever óbitos e não atingiu o estágio final de avaliação de impacto – nível I (quadro 2).

Listados abaixo, os estudos considerados mais significativos.

Ano da publicação	Nome do modelo ou autor	Tema	Observações	Estágio de desenvolvimento
1989	Parsonnet	Predição de mortalidade e morbidade	Parsonnet[36]: desenvolvido em 1989 para predição de óbitos e morbidades pós-cirurgia cardíaca incluindo a predição de sepse e IFO, submetido vários trabalhos de validação. Calculo complexo que inclui idade, cirurgia de emergência, função ventricular, características cirúrgicas, sexo, hemodiálise, pressão arterial, obesidade, gradiente átrio-ventricular, balão intra-aortico, aneurisma ventricular e estados cirúrgicos catastróficos.	Nível I
1990	New York	Predição de complicações e ajuste de risco em cirurgias cardíacas	Desenvolvido com dados de 30 hospitais americanos New York[35]–, utilizado como benchmark, e validado externamente por vários estudos. Inclui idade, sexo, doença coronariana com estenose importante, angina instável, função ventricular, infarto recente, balão intra-aortico, insuficiência cardíaca, diabetes, obesidade mórbida, DPOC, diálise e reoperação.	Nível II
1992	APACHE II	Construído para predição de mortalidade, validado para infecções pós-cirurgia cardíaca	Foi originalmente [16] desenhado e mais utilizado para predição de óbitos em UTI geral, foi validado por Kreuzer para prever sepse pós-cirurgia cardíaca, utilizando 19 como ponto de corte.	Nível IV
1992	Cleveland	Predição de óbitos e morbidade	Avalia mortalidade e morbidade[23] incluindo infecções em pacientes submetidos à revascularização do miocárdio (RVM). Utiliza originalmente 13 fatores preditivos, validado com algoritmo de nove fatores: idade, doença cerebrovascular, DPOC, anemia, insuficiência renal, função ventricular, insuficiência mitral, reoperação e cirurgia de emergência.	Nível IV
1995	Canadense	Predição de óbitos e tempo de permanência	Desenvolvido pelo Ministério da Saúde Canadense[20] para predição de mortalidade, duração de internação hospitalar total e em terapia intensiva, utiliza dados computadorizados avaliando seis fatores: idade, sexo, função ventricular, tipo de cirurgia e cirurgia de urgência. Validado em nova amostra.	Nível IV
1999	EUROSCORE	Óbitos em cirurgia cardíaca	O European System for Cardiac Operative Risk Evaluation [26] foi introduzido em 1999 e usa regressão logística para identificar e dar peso aos fatores associados à mortalidade intra-hospitalar após cirurgias cardíacas. Os modelos logístico e aditivo foram validados por Toumpoulis[27] para predição de mortalidade hospitalar, mortalidade de três meses, internação prolongada e complicações pós-operatórias incluindo infecções. Não foi validado para predição de acidente vascular cerebral, infarto do miocárdio, mediastinite, complicações gastrintestinais e sangramentos. É um dos escores mais utilizados no Brasil para ajuste de risco.	Nível IV

Quadro 2 – Modelos de predição em infecção e/ou cirurgia cardíaca (Continua).

Ano da publicação	nome do modelo ou autor	Tema	Observações	Estágio de desenvolvimento
1999	Staat	Predição de eventos adversos graves pós RVM	Tem como objetivo prever eventos graves [41] pós RVM, incluindo infecções. Inclui as seguintes variáveis: insuficiência cardíaca direita, arritmias, reoperação, DPOC, alterações no segmento ST, IMC. Sem validação.	Nível IV
2000	Escolano	Ocorrência de Infecção hospitalar e óbito em UTI	Modelo heterogêneo multi-estado semi-Markov [25] –, publicado em 2000, com o objetivo de prever infecções hospitalares em pacientes internados em UTI. O modelo define cinco estados: internação na UTI, primeira infecção simples, primeira infecção complicada, morte ou alta. Validado internamente por divisão da amostra.	Nível II
2001	Arozullah	Predição de pneumonia após cirurgias não-cardíacas	Desenvolvido para predizer pneumonia [17] em cirurgias não cardíacas, inclui tipo de cirurgia, idade, status funcional, perda de peso, DPOC, anestesia geral, alteração do sensório, AVC, uréia, transfusão, cirurgia de emergência, uso prolongado de esteróides e alcoolismo. Sem validação	Nível II
2002	Platt	Predição de IFO pós-cirurgia cardíaca por dados automatizados	Predição de infecção de ferida cirúrgica [37] – pós cirurgia de revascularizaçao do miocárdio, baseada em dados administrativos automatizados.	Nível IV
2002	Russo	Predição de IFO pós-cirurgia cardíaca	Desenvolvido para substituir o NNIS[40] –como parâmetro para estratificar pacientes conforme o risco de IFO pós-cirurgia cardíaca. Inclui obesidade, doença vascular periférica ou cerebral, diabetes, duração da cirurgia maior que cinco horas. Sem validação.	Nível IV
2003	Amphiascore	Eventos adversos pós-cirurgia cardíaca	Desenvolvido por Huijskes[15] para predição de óbitos, tempo de internação e MACE (eventos adversos cardiovasculares graves). Validado internamente por divisão da amostra.	Nível II
2003	Chong	Eventos adversos pós-cirurgia cardíaca por rede neural	Rede neural [22] –para predição de eventos adversos (morte, parada cardíaca, coma, insuficiência renal, AVC, reinfarto, ou ventilação mecânica prolongada) com 18 variáveis, submetida à validação cruzada.	Nível IV
2003	Dunning	Predição de ventilação prolongada pós- cirurgia cardíaca	Criado para predizer ventilação mecânica prolongada após cirurgia cardíaca [24] – inclui como variável o escore Parsonnet, fração de ejeção, diuréticos, uso de nitratos EV, idade, creatinina e reoperação. Submetido a validação externa em conjunto com Euroscore, sem demonstrar vantagens sobre este.	Nível III
2003	Kohli	Predição de infecção de esterno	Ciado com objetivo de predizer infecções de esterno i[31] – inclui reoperação, diabetes, internação maior que três dias na UTI, uso de artéria mamária. A validação mostrou uma AUC de 0,64. Submetido apenas à validação interna.	Nível IV

Quadro 2 – Modelos de predição em infecção e/ou cirurgia cardíaca (Continuação).

Ano da publicação	Nome do modelo ou autor	Tema	Observações	Estágio de desenvolvimento
2004	NNIS	Criado com objetivo e ajuste de risco para IFO pelo CDC	National Nosocomial Infections Surveillance [34]– desenvolvido pelo CDC para ajuste de taxas de infecção cirúrgica para várias categorias cirúrgicas, incluindo cirurgias cardiotorácicas. Prevê pontos conforme a duração da cirurgia, potencial de contaminação e risco conforme ASA (Sociedade Americana de Anestesiologia). Apesar de validado e de ser utilizado em vários centros no mundo, não é considerado um bom instrumento para cirúrgica cardíaca, uma vez que praticamente todas as cirurgias são limpas e com doença de base descompensada, resultando em pouca discriminação entre os extratos.	Nível IV
2004	Northern New England	Predição de óbitos em cirurgias valvares	Northern New England (NNE). Em 2004 foi derivado um modelo visando predizer o óbito em cirurgias valvares, incluindo idade, superfície corporal, reoperação, creatinina, AVC, fibrilação atrial, insuficiência cardíaca, cirurgia de urgência ou emergência e cirurgia combinada. Sem validação.	Nível III
2005	STS	Predição de IFO e de sepse	Society of Thoracic Surgeons [42]. Divide-se em STS pré-operatório e intraoperatório, com avaliação pré-operatória da presença de IMC, hemodiálise, choque cardiogênico, idade, imunodepressão, diabetes e adicionalmente de tempo de perfusão, balão intra- aortico e três ou mais anastomoses coronarianas. Validado em nova população por Paul[43].	Nível III
2005	Thakar	Predizer insuficiência renal pós-cirurgia cardíaca	Proposto para predizer insuficiência renal [44] pós cirurgia cardíaca, selecionando 13 preditores independentes em uma regressão logística. Submetido apenas a validação cruzada.	Nível II
2006	Ivanov	Eventos adversos em cirurgia cardíaca	Desenhado para predição de óbitos pós-cirurgia [30] – cardíaca e ajuste de risco para avaliação de desempenho dos cirurgiões. Sem validação. Também publicou em 2006 o escore de Toronto.	Nível II
2006	McGregor	Escore para aquisição de patógenos multirresistentes	Estudo visando desenvolver RPC para prever infecção hospitalar por bactérias multirresistentes [33]-. Inclui diabetes, úlcera péptica, doença respiratória, doença renal, câncer e pacientes transplantados. Validado internamente.	Nível IV
2006	TRS	Predição de eventosadversos pós-cirurgia cardíaca	Toronto Risk Score [45] – escore complexo com 18 variáveis, desenvolvidas para predizer eventos adversos pós-cirurgias cardíacas, incluindo infecções e aumento no tempo de internação. Não validado.	Nível II
2007	Barbini	Predizer complicações pós- cirurgia cardíaca por várias metodologias	Barbini[18] desenvolveu oito modelos, incluindo Bayes linear e quadrático, modelo de vizinhança k-nearest, regressão logística, Higgins, sistemas de escore direto, e modelos de rede neural com uma e duas camadas. Os modelos visavam predizer complicações cardiovasculares, respiratórias, neurológicas, infecciosas e hemorrágicas após cirurgias cardíacas. Os modelos foram validados apenas por divisão da amostra original.	Nível I

Quadro 2 – Modelos de predição em infecção e/ou cirurgia cardíaca (Continuação).

Ano da publicação	Nome do modelo ou autor	Tema	Observações	Estágio de desenvolvimento
2007	Chang	Predizer alta para asilo pós-cirurgia cardíaca	Construído para avaliar o risco de pacientes idosos serem internados em asilos após cirurgia cardíaca [21] – Inclui sexo, osteoartrose, insuficiência cardíaca, fibrilação atrial, DPOC, infarto do miocárdio, doença carotídea oclusiva, anemia, obesidade, e doença renal ou ureteral prévia. Submetido a validação interna.	Nível IV
2007	Friedman	Predição de IFO	Construído para predição de infecção de ferida operatória (IFO)[28]:) pós-cirurgia cardíaca, utiliza regressão logística e cria escore que adiciona um ponto para diabetes, um ponto para índice de massa corporal (IMC) entre 30 e 35 e dois pontos para IMC≥35. Cada ponto do escore representa aproximadamente o dobro do risco de infecção de ferida cirúrgica. Apenas validação retrospectiva.	Nível III
2007	Hsieh	Óbitos, morbidade e tempo de permanência em cirurgia cardíaca	Desenvolvido por regressão logística [29] – em população na Tailândia, submetido apenas a validação interna.	Nível IV
2007	Lapresta-Moros	Predição de pneumonia hospitalar	Desenhado para predição de pneumonia em UTI geral [32]-. Inclui no modelo nutrição parenteral e enteral, cateter nasogástrico, traqueostomia, ventilação mecânica cirurgia prévia, coma e diabetes. Submetido a validação interna.	Nível IV
2007	Ostrosky-Zeichner	Predição de candidíase invasiva	Estudo desenvolvido para prever a ocorrência de candidíase invasiva [38] Unidades de Terapia Intensiva, apresentando uma sensibilidade de 34%.	Nível II
2007	Reddy	VM prolongada	Modelo [39]- para prever que fatores pré-operatórios que predizem a necessidade de ventilação mecânica prolongada pós-cirurgia cardíaca, identificando idade, volume expiratório forçado <70%, creatinina, tabagismo, fração de ejeção, doença vascular periférica, infarto miocárdio recente, ventilação mecânica, cirurgia cardíaca prévia, cirurgia de urgência ou emergência, cirurgia mitral ou aórtica e uso de CEC.	Nível III
2008	CAHTA	Óbitos em cirurgia cardíaca	O modelo de risco Catalan Agency for Health Technology Assessment [19] foi derivado na Catalúnia em 1994 para cirurgias cardíacas de grande porte como valvares, coronarianas e combinadas. Validado e avaliado impacto.	Nível IV

Quadro 2 – Modelos de predição em infecção e/ou cirurgia cardíaca (Conclusão).

Considerando a potencial utilidade de modelos de predição para infecções em cirurgia cardíacas e que não existem na literatura modelos de predição em suficientemente avaliados ou submetidos à validação externa, o desenvolvimento de técnicas de avaliação de desempenho dos modelos, assim como a criação de um modelo específico para pneumonia pós-operatória, são lacunas que devem ser preenchidas justificando o objeto da tese.

1 **OBJETIVOS**

1.1 **Objetivo principal**

Revisar as técnicas e métodos utilizados para validar modelos de predição de pneumonia pós-cirurgias cardíacos aplicá-los no desenvolvimento de uma regra de predição clínica para pneumonia e avaliar seu desempenho por validação externa.

1.2 **Objetivos secundários**

- Avaliar criticamente técnicas e métodos de determinação do desempenho de modelos preditivos (Artigo I).
- Construir e validar internamente nomograma para predição de pneumonia pós-cirurgia cardíaca, baseado no modelo de regressão logística (Artigo II).
- Validar externamente e comparar modelos de árvore de classificação e regressão logística para predição de pneumonia pós-cirurgia cardíaca (Artigo III).

2 METODOLOGIA

2.1 Fluxo da pesquisa

Na figura 1 apresenta-se o fluxo utilizado na pesquisa. O trabalho original desenvolvido no mestrado incluiu para modelagem pacientes de um hospital público e de um privado em 2002. Para o doutorado apenas os pacientes do hospital privado (HPC) foram incluídos. Esta amostra de pacientes serviu como base para criação do modelo de regressão logística e do modelo de árvore de classificação (CART). Os dois modelos foram submetidos à validação interna. A partir do modelo logístico foi criado um nomograma. Os dois modelos foram posteriormente aplicados a uma nova amostra de um hospital público (INC) em 2006 (validação externa).

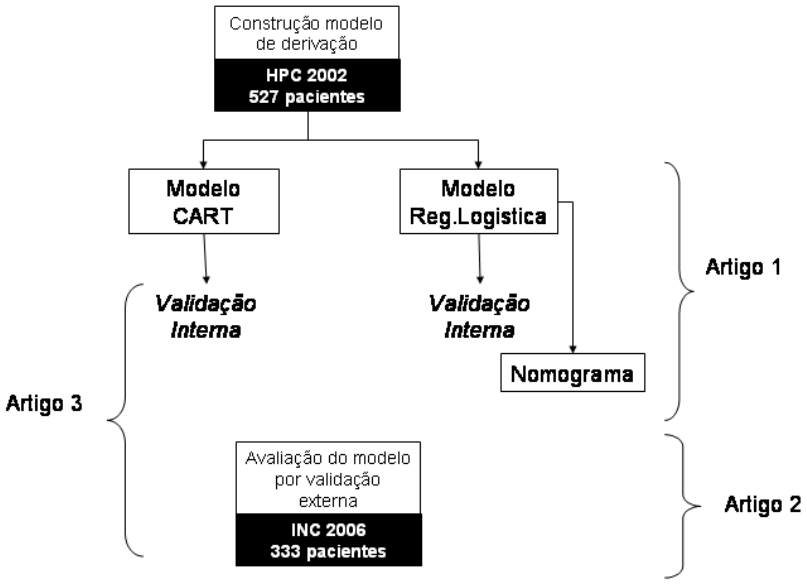

Figura 1 - Fluxo da Pesquisa

2.2 **Descrição das variáveis**

a) Critérios diagnósticos

A ocorrência de pneumonia foi diagnosticada por vigilância ativa realizada por médico infectologista, conforme os critérios descritos pelo CDC-EUA: novo infiltrado ao RX com escarro purulento e febre ou leucocitose até 10 dias de pós-operatório de cirurgia cardíaca. Não foi exigida a confirmação microbiológica.

b) Variáveis de exposição

Fez–se a opção de trabalhar apenas com variáveis pré-operatórias, a fim de desenvolver um modelo que permitisse estabelecer medidas preventivas antes da intervenção cirúrgica. Foi eleito para investigação o seguinte conjunto de variáveis:

Demográficas: sexo (masculino/feminino) e idade (em anos);

Antropométricas: peso (em kilogramas, mensurado na enfermaria ou informado), altura (em metros, mensurada na enfermaria ou informada), índice de massa corporal (IMC = peso/altura^2);

Doenças pré-existentes ou fatores reconhecidamente agravantes da cirurgia, definidas de acordo com os seguintes critérios:

Diabetes (sim/não): história de diabetes, glicemia de jejum >134mg% ou uso prévio de medicação hipoglicemiante;

Doença vascular periférica (sim/não): claudicação no exercício ou repouso, amputação por insuficiência arterial, doença aorto-ilíaca obstrutiva, cirurgia prévia de revascularização periférica, angioplastia ou *stent* periféricos, aneurisma de aorta abdominal, redução de pulso arterial ao exame clínico ou por método complementar;

Doença Pulmonar Obstrutiva Crônica - DPOC (sim/não): história de DPOC, diagnósticos através de exames complementares ou uso de corticóides ou broncodilatadores;

Tabagismo (sim/não);

Tempo de internação prévio (dias);

Creatinina sérica (mg%): último valor da dosagem antes da cirurgia; Diálise (sim/não).

c) Marcadores da gravidade da doença cardíaca:

Função ventricular (normal-leve/moderada-grave): avaliada através de ecocardiograma ou durante o cateterismo;

Escala clínica da New York Heart Association (NYHA):

I: Doença cardíaca que não promove restrições na atividade física.

II: Pequenas restrições, confortáveis em repouso. A atividade física promove fadiga, palpitações, dispnéia ou dor anginosa.

III: Restrições acentuadas à atividade física, confortável em repouso. Atividade menor que a normal promove fadiga, palpitações, dispnéia ou dor anginosa.

IV: Impede a realização de qualquer atividade física sem sentir desconforto. Os sintomas de insuficiência cardíaca ou de angina podem estar presentes no repouso.

Esta variável foi operacionalizada na forma dicotômica (disfunção até grau III ou grau IV).

Insuficiência mitral (sim/não): diagnosticada por ecocardiograma ou cineangiocoronariografia;

Angina instável (sim/não): presença de isquemia refratária progressiva requerendo hospitalização em terapia intensiva e medicação endovenosa para controle;

Tamanho do átrio esquerdo (em milímetros): obtido ao ecocardiograma;

Hipertensão pulmonar (sim/não): definida pelo ecocardiograma como pressão de artéria pulmonar superior a 35 mm Hg;

Infarto agudo do miocárdio nos últimos 3 meses - IAM 3 (sim/não): história comprovada até 3 meses passados de pelo menos 2 dos seguintes critérios:

-Dor torácica típica 20 minutos, não aliviada por repouso ou nitratos;

- Aumento enzimático (Creatino Kinase fração mb >5% da CK total, troponina >0,2 microgramas/ml);

-Novas mobilidades anormais da parede ao ecocardiograma;

- ECG seriado mostrando alterações na linha de base ou em ondas seriadas ST-T e/ou Q com 0,03 segundo de espessura e/ou mais de 1/3 do complexo QRS total em duas ou mais derivações contínuas.

Infarto agudo do miocárdio entre 3 e 6 meses passados - IAM 6 (sim/não): história referida ou obtida no prontuário de infarto entre 3 e 6 meses passados.

Arritmia (sim/não): Presença de arritmia nas duas semanas anteriores à cirurgia, por documentação clínica de um dos seguintes eventos: fibrilação/flutter atrial, bloqueio cardíaco, taquicardia ventricular sustentada ou fibrilação ventricular.

Variáveis cirúrgicas:

Cirurgia cardíaca prévia (sim/não);

Tipo de cirurgia:

-Eletiva: poderia ser protelada sem aumentar o risco de comprometimento do resultado para o coração;

-Urgência: procedimento necessário durante a hospitalização para minimizar as chances de deterioração clínica posterior, na ausência da condição eletiva e da condição de emergência;

-Emergência: presença de um dos abaixo:

-Disfunção isquêmica (isquemia progressiva, infarto agudo em curso nas 24 horas que antecedem a cirurgia ou edema agudo de pulmão que necessite entubação);

- Disfunção mecânica: choque com ou sem suporte circulatório.

Esta variável foi operacionalizada na forma dicotômica (cirurgia de emergência ou não).

2.3 **Detalhamento da construção dos modelos**

Inicialmente, avaliaram-se as diferenças clínicas e demográficas das populações dos dois hospitais envolvidos no estudo através de testes para diferenças de proporções, testes t de Student e testes de Mann-Whitney, utilizando-se o software Stata 9.0 (Stata Base Reference Manual, 2001). Foram utilizadas duas técnicas para identificar os fatores prognósticos mais importantes para ocorrência de pneumonia: regressão logística múltipla e árvore de classificação e regressão (CART).

A análise revelou que os pacientes submetidos a cirurgias de emergência apresentaram sete vezes mais chances de apresentar pneumonia no pós-operatório (Tabela 1).

Tabela 1 – Distribuição das variáveis de preditivas para pneumonia-análise univariada

Variável	***Valor p***	***Razão de chances***
Diabetes	0,18	1,59
Insuficiência mitral	0,01	2,79
Angina instável	0,57	1,20
Revascularizaçao prévia	0,44	1,53
Diálise	0,39	1,92
Doença vascular periférica	0,18	1,566
Tabagismo	0,78	1,14
Hipertensão pulmonar	0,03	2,88
Tamanho do átrio esquerdo	0,43	1,02
IAM <3 meses	0,01	2,49
IAM <6 meses	0,56	1,44
Ritmo não sinusal	0,02	2,66
Cirurgia emergência*	<0,01	7,09
Sexo	0,65	1,17
NYHA=4	<0,01	4,38
Disfunção ventricular moderada/grave	<0,01	3,74
Idade	<0,01	1,06
Dias de internação	0,618	0.91
IMC<=18.5	0,39	2,47
IMC>=30	0,14	0,42

*cirurgia de emergência comparada a cirurgia eletiva+ cirurgia de urgência

Foram incluídas no modelo de regressão logística as variáveis cuja associação com o desfecho, na análise bivariada apresentou significância estatística da ordem de até 20% ($p \leq 0,2$), permanecendo no modelo final aquelas cuja associação com o desfecho apresentou nível de significância de 5%.

A árvore de classificação e regressão (CART):

Oferece uma alternativa para os modelos de regressão logística para construção de modelos preditivos. Foi construída com a utilização do software R.

Histórico

Inicialmente inspirado no AID (Automatic Interation Detection), sua metodologia foi descrita e formalizada no livro *Classification and Regression Trees* [15]. Foi utilizada, inicialmente, em aplicações não - médicas como reconhecimento de barcos por radar, conteúdo de cloro por espectro de massa, milhagem e peso de automóveis. Posteriormente seu espectro de aplicações se ampliou para a análise de poluentes, prognóstico após infarto do miocárdio e diagnóstico de câncer.

Estrutura

O modelo baseado em árvore é uma técnica de análise exploratória para revelar a estrutura dos dados, sendo útil quando existe um grupo de variáveis preditivas e apenas uma variável de desfecho. Para desfechos binários, CART constrói um sistema de classificação binário (presença ou ausência do desfecho em questão) através de partições sucessivas, dividindo os dados em subgrupos mais homogêneos a cada divisão. A cada divisão (“nó”) o algoritmo seleciona a variável com maior capacidade de discriminação entre dois desfechos (no caso, ocorrência ou não de pneumonia). Seguindo uma estrutura hierárquica, a primeira divisão corresponde à variável com maior poder discriminatório e, a partir daí, as divisões subseqüentes seguem o mesmo critério. Os “ramos” são acrescentados à “árvore” pelo algoritmo até que se obtenha um grupo mais homogêneo em termos da probabilidade de apresentar o desfecho em questão ou contenha poucas observações (“folhas”).

A CART cria uma grande árvore com um mínimo de erros de classificação, porém muito “ajustada” aos dados.

Poda e "Encolhimento” (*prune/ shrinking)*

O ajuste excessivo aos dados pode ser corrigido pelas técnicas de poda e

encolhimento. Na poda, as probabilidades estimadas são semelhantes às da árvore original, fornecendo uma descrição mais sucinta dos dados.

No encolhimento, as probabilidades são totalmente diferentes, mas o erro de classificação é menor do que na árvore original. A escolha depende da opção entre simplicidade versus acurácia.

No estudo em questão foi utilizada a redução do tamanho (“poda”).

Escolha da árvore ideal

A escolha da melhor árvore foi realizada a partir da inspeção de gráfico que representa os ganhos em redução dos erros de classificação em função do tamanho da árvore (figura 2). Como no gráfico em questão metade da redução no desvio médio é explicada por 20 ramos, optou-se por trabalhar com três tamanhos de árvores: 12, 17 e 20 ramos. Foram comparadas as áreas sob a curva ROC produzidas pelas árvores de diferentes tamanhos com a utilização do software Stata 7.0. A probabilidade de pneumonia prevista de cada indivíduo na coorte foi estimada a partir de cada modelo e então comparada com o desfecho real. Foi utilizada a área sob a curva ROC (AUC), sensibilidade, especificidade, valores preditivos positivos e negativos e acurácia como critérios para avaliar o desempenho dos modelos.

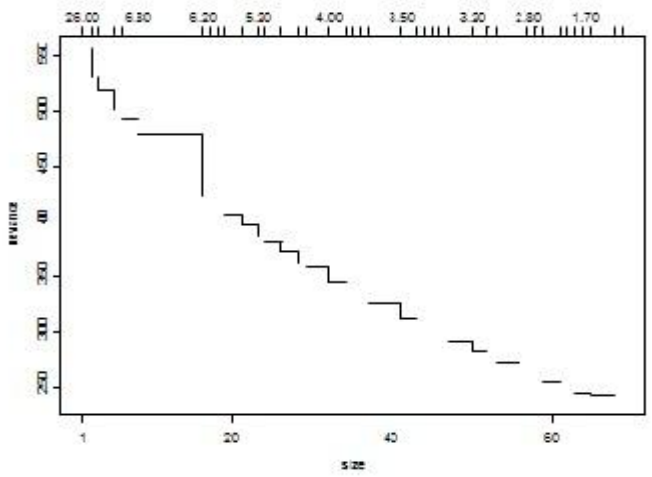

Figura 2 – Gráfico de redução no desvio médio conforme número de ramos

3 ARTIGO I - A AVALIAÇÃO DE DESEMPENHO DE MODELOS PREDITIVOS NÃO SE LIMITA À ÁREA SOB A CURVA ROC

Artigo I. Resumo

Objetivo: Fazer uma revisão dos métodos e medidas de avaliação do desempenho de modelos preditivos.
Metodologia: Foram examinados os métodos e medidas de avaliação de desempenho dos modelos preditivos encontrados na literatura da área. Estas técnicas foram classificadas em relação aos níveis de classificação das evidências para estudos envolvendo modelos preditivos. Freqüentemente este processo de avaliação objetiva o aperfeiçoamento dessas regras de predição. Os métodos examinados foram: (1) avaliação do modelo de derivação; (2) validação interna; (3) validação externa e (4) análise de impacto. As medidas são classificadas como relativas à discriminação (índice C, curva ROC, ângulo de discriminação, gráfico de caixas), a calibração (calibração em larga escala, ângulo de calibração, teste de Hosmer-Lemeshow e de Goeman-Le Cessie) e globais (R^2, escore de Brier). São apresentadas análises críticas dos métodos e das medidas utilizadas para avaliação de desempenho. Como exemplos foram aplicados estes métodos em um modelo de predição de pneumonia em pacientes pós-cirurgia cardíaca para demonstrar empiricamente as propriedades destas técnicas e medidas.
Resultados: A escolha do método de avaliação depende do nível de evidência pretendido, enquanto a escolha das medidas depende do aspecto prioritário da avaliação, se discriminação ou calibração. Os objetivos do modelo (diagnóstico ou prognóstico) impactam tanto nos métodos quanto nas medidas de avaliação. No exemplo do modelo de predição da pneumonia, a aplicação do método validação externa resultou em redução dos valores das medidas de discriminação e calibração quando comparados com os demais métodos. Estes resultados indicam a importância de métodos mais robustos para a avaliação dos modelos de predição. Nas três técnicas usadas, as medidas de calibração apresentaram melhor desempenho do que aquelas de discriminação.
Conclusões: Existem marcadas diferenças das técnicas de avaliação de desempenho dos modelos de predição. Provavelmente as ferramentas usadas na vigilância epidemiológica e na prática médica decorrem de modelos preditivos insuficientemente testados. Devem ser mais bem divulgadas as metodologias empregadas na análise crítica das regras de predição clínica e das medidas disponíveis para a apropriada avaliação do seu desempenho.

Artigo II. Introdução

Modelos de predição são úteis para varias aplicações médicas e não-médicas. Dentre as aplicações não-médicas destacam-se os modelos para previsão de tempestades e para sinistros em seguros de automóveis. Na área da saúde pública, inclui a vigilância epidemiológica de doenças como a Encefalite do Oeste do Nilo[1], detecção de epidemia de malária, febre do Vale Rift e dengue, diagnóstico de condições de meningite [2] até diversas formas de prognóstico em pacientes com neoplasia[3]. A modelagem destes modelos de predição parte habitualmente de uma série de informações simples de várias fontes e a partir de um método estatístico se estabelece a probabilidade de um desfecho.

A prática clínica envolve em suas decisões diárias uma avaliação de probabilidade de diagnóstico e de prognóstico, baseadas em dados de anamnese e de exame físico, radiológicos e laboratoriais. As regras de predição ou modelos preditivos visam aumentar a acurácia do diagnóstico ou do prognóstico clinico[4]. Além deste uso mais simples, os modelos podem ser utilizados em ensaios clínicos randomizados para decidir sobre a inclusão de um paciente ou sobre o ajuste por uma covariada. Em estudos observacionais, a utilização recai em ajuste de fatores de confusão, ajuste do espectro de doença (“case-mix”) entre unidades diferentes[5], ou em uma mesma unidade comparação de uma incidência entre períodos temporais distintos. Um uso importante dos modelos é informar ao paciente sobre o seu prognóstico, em especial, em caso de neoplasias. Alguns modelos visam mais diretamente um suporte na decisão sobre o emprego de teste diagnóstico ou tratamentos de alto custo ou invasivos.

Alguns entusiastas propõem objetivos mais ambiciosos como os membros do *Evidence-Based Medicine Working Group*, que defendem que os modelos podem modificar o comportamento clínico e reduzir custos desnecessários, mantendo a qualidade do cuidado e satisfação do tratamento[6]. Muitos modelos têm sido mais utilizados para avaliação de prognóstico em terapia intensiva, em especial, para predizer o número de óbitos, sendo tomados como índice de mortalidade ajustada, como no caso do modelo STS para pós-operatório de cirurgia cardíaca[7].

Quando um modelo é desenvolvido, recomenda-se que o seu desempenho seja quantificado e comparado com o de outros modelos quando disponíveis. A avaliação de desempenho dos modelos preditivos tem sido negligenciada pelos pesquisadores que elaboram as regras de decisão aplicadas à clínica e à vigilância epidemiológica. Na literatura são escassos os textos que sistematizam as técnicas de avaliação dos modelos.

Artigo III. Níveis de evidência para incorporação de regras de predição clínica

O desenvolvimento completo de um modelo preditivo envolve três etapas essenciais e, antes que o modelo seja incorporado à rotina clínica, estas etapas se relacionam à produção de níveis de evidência variando do nível IV (apenas derivação) até I (todas as etapas cumpridas), conforme proposto por McGinn e Guyatt[8] e adaptado na figura 9.

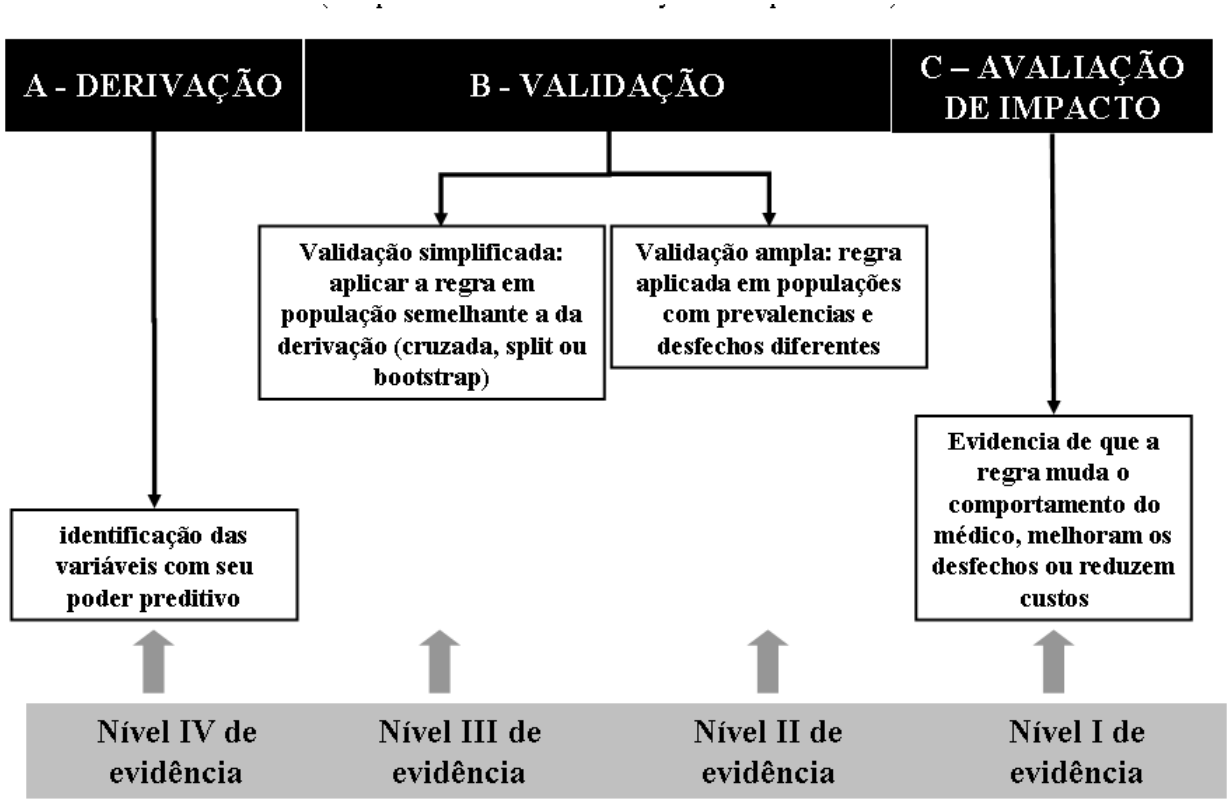

Figura 1 – Etapas de desenvolvimento de um modelo preditivo ou regra de predição clínica (Adaptado de McGinn e Guyatt com permissão)

O processo completo de validação necessita de vários estudos em diferentes centros para que se possa testar adequadamente a acurácia da regra. A maioria dos estudos desenvolve processos denominados de "derivação e validação" em um cenário onde a validação é limitada à utilização de técnicas estatísticas em uma mesma amostra, incluindo divisão da amostra ou técnicas de reamostragem, devendo ser considerada apenas como validação interna. Modelos preditivos não-validados não estão prontos para utilização, embora possam ter outras utilidades como determinar a importância de cada um dos preditores na ocorrência do desfecho e identificar outros fatores que não mostraram associação forte com o desfecho[4]. Por exemplo, no modelo de predição de óbito em paciente com pneumonia descrito por Fine[9] a leucocitose não impactou no prognóstico dos pacientes, levando os clínicos a valorizarem menos este dado laboratorial.

A regra não validada pode ser falha por três motivos básicos: (a) associações ao acaso, (b) os preditores podem ser peculiares apenas naquele grupo de pacientes e (c) a regra pode ser de utilização muito complexa ou de difícil aplicabilidade na prática. Existem algumas

poucas exceções ao raciocínio proposto, como no caso do modelo para predição de óbitos em pacientes com Síndrome Coronariana Aguda[10], em cujo estudo foi incluído na derivação e na coorte de validação um grande número de centros e um volume suficiente de pacientes para garantir a capacidade de fazer inferências com base nos resultados, embora a questão da avaliação de impacto não tenha sido solucionada.

A validação pressupõe que a aplicação repetida da regra leva a resultados semelhantes, com populações distintas da originalmente testada. Quanto mais heterogêneas forem as amostras de pacientes em que o modelo foi testado, maior será sua capacidade de generalização[11].

(a) Métodos de avaliação de desempenho das regras de predição clínica

a - Avaliação do modelo de derivação

Além dos níveis de evidência, alguns pontos metodológicos são fundamentais na análise crítica de um estudo sobre modelos preditivos[12]. Os pacientes devem ser escolhidos sem viés e representando um espectro variado de gravidade. Outro aspecto importante é que os critérios de definição dos preditores sejam padronizados e que o pesquisador seja cego em relação à ocorrência do desfecho. No exemplo da pneumonia, o conhecimento do diagnóstico final pode influenciar o radiologista na detecção de imagens de consolidação. Por outro lado, quem determina se o desfecho é presente ou não também não deve conhecer previamente se o paciente era portador de condições de alto risco.

A derivação consiste na avaliação do modelo construído, tendo dois pontos fundamentais: desenho do estudo que origina os dados e o tipo de modelagem escolhida. Os estudos que vão gerar a informação a ser modelada poderão depender do objetivo da predição, podendo ser prognóstico, diagnóstico, desfechos contínuos, binários, etc. Os modelos mais utilizados são as coortes, séries temporais, casos-controle e transversais.

O tipo de modelagem também é vinculado ao desfecho desejado, sendo a regressão logística utilizada com mais freqüência pela presença de desfechos binários. Outros métodos de modelagem incluem árvores de regressão e classificação, redes neurais, regressões lineares, modelos generalizados aditivos, modelos bayesianos, regressão de Cox e análise de Kaplan-Meier.

b- Validação interna

O objetivo de um modelo é predizer um desfecho quando aplicado a uma nova amostra já que os dados da amostra inicial servem apenas para o aprendizado futuro. É essencial que seja documentada a reprodutibilidade do modelo (validade interna) como passo inicial, isto é, se o modelo proposto é válido dentro da amostra onde este se originou, visando criar um modelo sem superajuste. São métodos estatísticos que possibilitam a testagem sem que seja colhida uma nova amostra. As técnicas mais utilizadas são a divisão da amostra (*"split*"), a validação aparente, validação cruzada, e técnicas de reamostragem como o "*bootstrap*".

c- Validação externa

Consiste na testagem na transportabilidade ou capacidade de generalização do modelo. Pode ser geográfica, temporal ou totalmente independente, devendo aproximar-se do manejo utilizado para testar toda hipótese científica[13] por testes repetidos comprovando a idéia original. Os motivos mais freqüentes de falha na validação são: o tamanho da amostra original, técnicas de análise estatísticas falhas, diferenças na definição dos preditores ou do desfecho, método de cálculo da predição e a dificuldade de obtenção dos preditores.

d- Avaliação de impacto

A avaliação o impacto abrange diversos aspectos, desde a aplicabilidade da regra (tempo para cálculo, exames necessários, aceitação pelo corpo clínico) até a quantificação de efetividade clínica da utilização da regra, que pode incluir desde agilidade no atendimento em um setor de emergência, redução nos erros diagnósticos até pontos importantes como a redução de óbitos ou economia de recursos (custo-efetividade).

Idealmente, estes estudos devem incluir randomização individual ou em blocos (por exemplo, randomizar diversas UTIs) quando a estratégia individual levar a "contaminação" na conduta médica. Quando a randomização não é possível, podem ser utilizados estudos quasi-experimentais e até avaliações antes-e-depois[14]. Estratégias para facilitar a utilização da regra podem ser desenvolvidas (cartões de bolso, lembretes eletrônicos, programas para computadores de bolso), e o cenário deve ser o mais próximo possível da realidade (ensaios pragmáticos)[15].

A translação do modelo, ou seja, a incorporação à prática clínica enfrenta desafios em especial a dificuldade na utilização da regra e a crença do médico de que as regras "engessam" a prática e não acrescentam nada a sua própria *expertise*[16].Mesmo uma regra metodologicamente adequada e testada externamente com bons resultados pode não resultar em mudança na prática clínica ou na redução de desfechos. Algumas regras tornam os estudos de impacto uma precondição para o seu uso como, por exemplo, a regra PORT para admissão de pacientes com pneumonia[4] em que foi realizado estudo de antes-e-depois.

O propósito do modelo deve ser considerado para escolha das técnicas de avaliação. [17]. Os modelos que visam o diagnóstico devem primariamente classificar adequadamente os indivíduos em dois grupos com e sem doença (discriminação) e tradicionalmente são avaliados pela sensibilidade, especificidade, valor preditivo positivo e negativo e pela razão de verossimilhança. Já os modelos prognósticos visam estimar riscos e classificar o individuo em estratos segundo as probabilidades de adoecimento, devendo primariamente ser avaliados pelos testes de calibração.

A avaliação do desempenho de um modelo pode ser gráfica ou empírica, envolvendo três tipos de medidas: discriminação, calibração e globais.

(b) Medidas de avaliação de desempenho das regras de predição clínica

Medidas de discriminação

Entende-se por discriminação a capacidade de separar os indivíduos em grupos com ou sem o desfecho[13]. Utilizando o exemplo, um modelo com boa capacidade discriminativa tem facilidade para separar grupos com e sem pneumonia no pós-operatório de cirurgia cardíaca. As medidas gráficas utilizadas na discriminação são a curva ROC e o gráfico de caixas de discriminação.

A curva ROC é a avaliação de desempenho mais conhecida na prática clinica, sendo utilizada cm muita frequência em exames laboratoriais, plotando-se em cuja construção consecutivos pontos de corte para a probabilidade predita de um desfecho binário, iniciando com 0% (todo individuo é classificado como positivo para o pneumonia), resultando em uma sensibilidade de 100% e uma especificidade de 0%. O ponto de corte da curva ROC mais

próximas do canto superior esquerdo apresenta melhor discriminação. A área sob a curva ROC (AUC) também é conhecida como estatística de concordância ou estatística C, descrita a seguir.

A estatística C é uma estatística ordenada por ranques para predições contra desfechos reais, relacionada à estatística D de Sommer, sendo insensível a erros de calibração como diferenças na média do desfecho[13]. A estatística C é interpretada no exemplo como a probabilidade de que o modelo prediga uma maior probabilidade de desfecho positivo para um paciente com pneumonia em comparação com um paciente sem pneumonia, escolhido aleatoriamente[13]. A estatística C varia de modelos não-informativos com um valor de 0,5 a modelos perfeitos com valor de um. Em modelos preditivos em medicina a AUC geralmente apresenta valores em torno de 0,8 pela dificuldade de predição em função da complexidade dos fenômenos[18]. A interpretação dos resultados habitualmente é interpretada como: [19] não-informativo 0,50 to 0,75; bom 0,75 a 0,92; muito bom 0,92 a 0,97; excelente 0,97 a 1,00.

O gráfico de caixas de discriminação é uma das formas de facilitar a visualização da capacidade discriminativa de um modelo preditivo (figura 10) representando uma medida de fácil entendimento e de comparação entre modelos diferentes. Sua elaboração inclui duas caixas distintas (pacientes com e sem desfechos), em que a amplitude dos dados é representada por linhas, cujo comprimento vai da caixa até o extremo dos dados. Os extremos correspondem aos quartis inferior (percentil 25) e superior (percentil 75) dos dados e à linha central da caixa a mediana. A altura da caixa é igual à amplitude interquartílica. Quanto menor a sobreposição das caixas, melhor a discriminação do modelo. A desvantagem é que a incidência do desfecho vai impactar na visualização das caixas, determinando que em pequenas incidências exista um menor impacto visual com “achatamento das caixas”. A tradução numérica do gráfico de discriminação é o angulo de discriminação. Também é afetada por baixas incidências do desfecho.

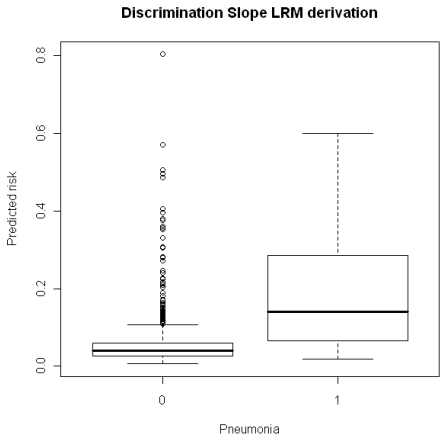

Figura 2 - Gráfico de discriminação para modelo de predição de pneumonia

O **ângulo de discriminação** é representado pela diferença absoluta entre a predição média dos indivíduos com e sem o desfecho, e quanto maior o valor, melhor a capacidade discriminativa.

Seção 3.02 Medidas de calibração

As medidas de calibração objetivam avaliar a capacidade de concordância entre o predito e o observado, em desfechos binários esta concordância é observada em grupos.

A calibração é representada graficamente pelo **gráfico de calibração** (figura 3) onde as probabilidades preditas são plotadas no eixo x o desfecho observado no y. Em um eixo central a 45º é representada a calibração perfeita, com algumas variações técnicas na construção dos pontos [13]. Para regressão linear, o gráfico de calibração é um gráfico simples de dispersão. Em desfechos binários a plotagem consistiria apenas zero e um no eixo dos Y, podendo ser substituído por técnicas de alisamento para estimar as probabilidades observadas do desfecho (p (y - 1)) em relação às probabilidades preditas, usando, por

exemplo, um algoritmo Loess[5]. Outra possibilidade é plotar indivíduos com probabilidades semelhantes e comparar a média da probabilidade predita e a média da observada representando uma apresentação gráfica do teste de Hosmer-Lemeshow, este agrupamento, embora comum, é considerado arbitrário e impreciso.

O **gráfico de calibração** pode ser descrito por um intercepto e por um ângulo. O intercepto indica a extensão do quanto às predições são muito baixas ou muito elevadas, denominado **Calibração em larga** escala. Na equação de regressão criada durante a derivação de um modelo logístico, a média das predições corresponde à média dos desfechos média (pneumonia predita) = média (pneumonia observada), relação garantida pelo coeficiente de regressão. Quando o modelo é aplicado a novo grupo de pacientes durante a validação externa a diferença observada entre a média (pneumonia predita pelo modelo) = média (pneumonia observada na validação) é conhecida como calibração em larga escala, tendo valores ideais iguais à zero. A força do efeito preditor é traduzida pelo ângulo de calibração com valores ideais iguais a um. O valor do **ângulo de calibração** pode ser utilizado com fator de recalibração do modelo para usos futuros uma vez que reflete dois aspectos, o super ajuste e diferenças reais na força dos preditores.

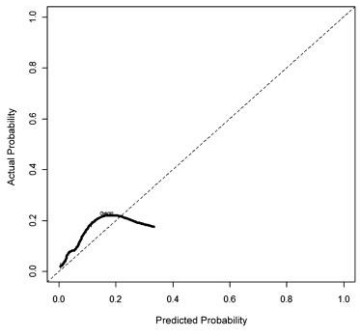

Figura 3 - Gráfico de calibração avaliando a calibração ideal (reta pontilhada a 45º), a probabilidade real(eixoY) e a probabilidade predita (eixo X)

Do ponto de vista estatístico a distância entre o valor predito e o observado é fundamental para quantificar o desempenho do modelo. Estas distâncias são relacionadas ao conceito de "adequação do ajuste" ("*goodness-of-fit"*) com modelos mais adequados, tendo menores distâncias. A calibração também pode ser avaliada pelos testes de ajuste do modelo de Hosmer-Lemeshow e Goeman-le-Cessie. Os dois testam a hipótese nula de que a

calibração é adequada, logo, tem valores ideais de p acima de 0,05. Hosmer-Lemeshow é o teste mais utilizado para desfechos binários, realizado pelo agrupamento dos pacientes em grupos de decis de probabilidade predita. A soma das probabilidades preditas de pneumonia é o número de desfechos esperados para aquele grupo. Este número esperado de pneumonias é comparado com o número real de pneumonias observados naquele grupo com um teste chi-quadrado. Em validação externa o teste deve ter 9 graus de liberdade. O teste tem muitas limitações como qual a melhor estratégia para separar os grupos e a pequena capacidade de detectar superajuste ou diferenças sistemáticas[20]. Já o teste de Goeman-le-Cessie tem capacidade de detectar efeitos de interação e não-linearidade não detectados no modelo pela análise dos resíduos[21] e este como hipótese alternativa de que os resíduos estão próximos no espaço e se desviam do modelo na mesma direção.

Seção 3.03 Medidas globais

A variabilidade explicada pelo modelo é denominada R^2, sendo utilizado com frequência em desfechos contínuos. Em modelos binários, o R^2 de Nagelkerke é o mais utilizado[5]. O R^2 corresponde a uma regra de escore logarítmica.

O escore de Brier é uma regra de escore quadrática, onde a diferença quadrática entre o desfecho binário real Y e as predições p são calculadas $(Y - p)^2$. O escore varia de zero (perfeito) a 0,25 (não informativo com 50% de incidência de desfecho). Quando a incidência do desfecho é pequena o escore máximo é menor. O escore varia de zero onde as predições são perfeitas a 0,25 para modelos não-informativos. Quando a incidência do desfecho é pequena, o escore máximo é menor.

Medida	Cálculo	Prós	Contras	Método
Medidas Globais				
R^2	Logaritmo das predições comparado com o desfecho real. Ideal: 100%	Fácil interpretação (variação explicada pelos fatores de risco)	Penalização importante para predições extremas com desfechos discordantes. Modelos diagnósticos têm R^2 maior que os prognósticos.	Avaliação do modelo derivação - Validação interna - Validação externa
Escore de Brier	Diferença quadrática entre os desfechos reais e os preditos $(Y-\hat{Y})^2$ Ideal: 0	Pode ser usado por qualquer tipo de desfecho (binário, continuo, censurado)	fetado pela prevalência pequena	valiação do modelo derivação - Validação nterna - Validação xterna
Medidas de Discriminação				
Estatística de Concordância (índice C)	Estatística com ordenação hierárquica para predições. A probabilidade de um indivíduo ao acaso com o desfecho ter uma predição maior do que um individuo sem o desfecho. Corresponde à área sob a curva ROC (AUC) Ideal: 1	Uso clínico amplo. Interpretação visual pelo gráfico da curva ROC. Robusta a variações da prevalência do desfecho.	Não é sensível a erros de calibração, pode variar pouco com inclusão de novos marcadores.	Avaliação do modelo derivação - Validação interna - Validação externa
Ângulo de Discriminação	Diferença absoluta na média das predições para pacientes com e sem o desfecho. Ideal: 1	Interpretação visual com gráfico de caixas	Afetado pela incidência baixa do desfecho	Avaliação do modelo derivação -Validação interna - Validação externa
Medidas de Calibração				
Calibração em larga escala	Diferença entre a média dos preditos ($\hat{Y}$) e a média dos desfechos (Ynovo) quando um modelo sofre validação externa. Ideal: 0	Visualizar calibração Ponto fundamental na validação externa	Interpretação difícil Útil apenas na validação externa	Validação interna - Validação externa
Ângulo de calibração	Angulo de regressão de um preditor linear Ideal: 1	Fator de "encolhimento" para ajuste Reflete o superajuste e diferenças no efeito dos preditores	Interpretação difícil Útil apenas na validação externa	Validação interna - Validação externa
Teste de Hosmer-Lemeshow (HL)	Compara observados versus preditos em grupos de pacientes (teste qui-quadrado) Ideal $p>0,05$	Conceitualmente fácil	Poder pequeno em pequenas amostras Falha para detectar super ajuste do efeito dos preditores. Pode falhar em modelos complexos.	Avaliação do modelo derivação - Validação externa
Teste deGoeman le Cessie	Correlação entre resíduos. Soma quadrática dos residuos atenuados. Ideal: $p>0,05$	Avalia não linearidade e efeitos de interação	Restrito a modelos logísticos ou multinomiais Interpretação difícil	Avaliação do modelo derivação-Validação externa

Quadro 1 - Prós e contras das medidas de avaliação de desempenho

(a) Demonstração empírica dos métodos de avaliação de desempenho

Estudo prospectivo tendo como objetivo a predição de pneumonia em pacientes submetidos à cirurgia cardíaca de grande porte foi submetido em 5 de março para publicação em periódico da saúde. Foram incluídos na derivação 527 pacientes atendidos em um hospital privado, sendo o modelo construído por regressão logística. Os dados foram submetidos à validação interna por técnica de reamostragem por bootstrap. O modelo foi submetido à validação externa com 333 pacientes de um hospital público, tendo sido avaliado seu desempenho. Os resultados demonstram na validação externa em relação à validação interna, uma perda na discriminação (passa de bom para não-informativo) e uma perda pequena na calibração, ainda que a calibração seja considerada adequada pelos testes de qualidade do ajuste do modelo, sugerindo uma possibilidade de recalibração pelo valor observado na calibração em larga escala.

O método de avaliação baseado apenas no modelo de derivação tende a superestimar tanto a discriminação e a calibração. O uso de técnicas de validação interna possibilita uma avaliação mais refinada. O desempenho de modelo mais robusto parece ser o da validação externa.

Artigo IV. Considerações finais

Um modelo preciso de uma análise crítica antes da sua incorporação na prática clínica que inclua minimamente validação externa como forma de garantir sua capacidade de generalização.

A análise tradicional limitada à curva ROC embora útil para classificação dos modelos, não pode ser a única estratégia para avaliação de desempenho. O desempenho adequado de um modelo deve contemplar a discriminação e a calibração, principalmente para modelos prognósticos[17]. A estatística C tem limitações importantes principalmente em relação à comparação entre diferentes modelos, *Peek*[22] estimou que amostras de até 5000 pacientes podem ser necessárias para detectar diferenças entre AUC.

Uma calibração e discriminação perfeitas são inconsistentes com modelos preditivos clínicos [18] que têm AUC habitualmente em torno de 0,8. A estatística C máxima depende da localização e da distribuição do risco na população [23].

Uma boa e simples estratégia para avaliar a capacidade do modelo é o exame do gráfico de calibração. Uma estratégia mais ampla e rigorosa de avaliação de desempenho do modelo deve incluir medidas globais (R2 e escore de *Brier*), discriminação (estatística C, ângulo de discriminação, gráfico de caixas, curva ROC) e de calibração (gráfico de calibração, calibração em larga escala, ângulo de calibração, testes de *Hosmer-Lemeshow* ou *Goeman-Le-Cessie*).

O desafio dos modelos preditivos é o balanço entre a simplicidade e reprodutibilidade de um modelo, sendo de um lado conveniente para o usuário de outro robusto o suficiente para ter um adequado poder preditivo. As regras de predição devem ser aliadas a um bom julgamento clínico e não um substituto estatístico para o bom senso no julgamento à beira do leito.

Artigo V. Referências

[1] Pan L, Qin L, Yang SX, Shuai J. A neural network-based method for risk factor analysis of West Nile virus. Risk Anal. 2008 Apr;28(2):487-96.

[2] Spanos A, Harrell FE, Jr., Durack DT. Differential diagnosis of acute meningitis. An analysis of the predictive value of initial observations. Jama. 1989 Nov 17;262(19):2700-7.

[3] Ross PL, Gerigk C, Gonen M, Yossepowitch O, Cagiannos I, Sogani PC, et al. Comparisons of nomograms and urologists' predictions in prostate cancer. Seminars in urologic oncology. 2002 May;20(2):82-8.

[4] Guyatt G. Users' guides to the medical literature : essentials of evidence-based clinical practice. 2nd ed. New York: McGraw-Hill Medical 2008.

[5] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology (Cambridge, Mass. Jan;21(1):128-38.

[6] McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-Based Medicine Working Group. Jama. 2000 Jul 5;284(1):79-84.

[7] Shroyer AL, Plomondon ME, Grover FL, Edwards FH. The 1996 coronary artery bypass risk model: the Society of Thoracic Surgeons Adult Cardiac National Database. The Annals of thoracic surgery. 1999 Apr;67(4):1205-8.

[8] McGinn T, P. W, Wisnisvesky J, Devreaux P, Stiell I, Richardson S. Clinical Prediction Rules. In: Guyatt G, ed. *Users' guides to the medical literature : a manual for evidence-based clinical practice*. 2nd ed. New York: McGraw-Hill Medical 2008:xxiii, 836 p.

[9] Fine MJ, Auble TE, Yealy DM, Hanusa BH, Weissfeld LA, Singer DE, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. The New England journal of medicine. 1997 Jan 23;336(4):243-50.

[10] Eagle KA, Lim MJ, Dabbous OH, Pieper KS, Goldberg RJ, Van de Werf F, et al. A validated prediction model for all forms of acute coronary syndrome: estimating the risk of 6-month postdischarge death in an international registry. Jama. 2004 Jun 9;291(22):2727-33.

[11] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Annals of internal medicine. 1999 Mar 16;130(6):515-24.

[12] Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. Jama. 1997 Feb 12;277(6):488-94.

[13] Steyerberg EW. Clinical prediction models : a practical approach to development, validation, and updating. New York ; London: Springer 2009.

[14] Igarashi A, Ikegami N, Yamada Y, Yamamoto-Mitani N. Effect of the Japanese preventive-care version of the Minimum Data Set--Home Care on the health-related behaviors of community-dwelling, frail older adults and skills of preventive-care managers: a quasi-experimental study conducted in Japan. Geriatrics & gerontology international. 2009 Sep;9(3):310-9.

[15] Skillgate E, Bohman T, Holm LW, Vingard E, Alfredsson L. The long-term effects of naprapathic manual therapy on back and neck pain - Results from a pragmatic randomized controlled trial. BMC musculoskeletal disorders. Feb 5;11(1):26.

[16] Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Annals of internal medicine. 2006 Feb 7;144(3):201-9.

[17] Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clinical chemistry. 2008 Jan;54(1):17-23.

[18] Diamond GA. What price perfection? Calibration and discrimination of clinical prediction models. Journal of clinical epidemiology. 1992 Jan;45(1):85-9.

[19] Swets JA. Measuring the accuracy of diagnostic systems. Science (New York, NY. 1988 Jun 3;240(4857):1285-93.

[20] Hosmer DW, Taber S, Lemeshow S. The importance of assessing the fit of logistic regression models: a case study. American journal of public health. 1991 Dec;81(12):1630-5.

[21] Goeman JJ, le Cessie S. A goodness-of-fit test for multinomial logistic regression. Biometrics. 2006 Dec;62(4):980-5.

[22] Peek N, Arts DG, Bosman RJ, van der Voort PH, de Keizer NF. External validation of prognostic models for critically ill patients required substantial sample sizes. Journal of clinical epidemiology. 2007 May;60(5):491-501.

[23] Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. Biostatistics (Oxford, England). 2005 Apr;6(2):227-39.

4 II - PREDIÇÃO PRÉ-OPERATÓRIA DE PNEUMONIA PÓS CIRURGIA CARDÍACA COM UM NOMOGRAMA

Resumo

Predizer o risco de infecções relacionadas a cirurgia cardíaca pode fornecer informações importantes para os pacientes e os cirurgiões. O objetivo do presente estudo foi desenvolver uma regra de predição para uma aplicação em nomograma baseado em fatores de risco pré-operatório para a ocorrência de pneumonia após cirurgias cardíacas.

Metodos. Dados de uma série de 527 pacientes, consecutivamente submetidos a cirurgias cardíacas entre Junho de 2000 e agosto de 2002, foram utilizados para desenvolver um modelo de prognóstico.O modelo foi validado internamente com um método de bootstrap. Análise de regressão logística múltipla foi realizada para identificar fatores preditivos para a ocorrência de pneumonia.Diversos fatores de risco simples e convencionais pré-operatórios foram avaliados. Um nomograma foi desenvolvido para melhorar a aplicabilidade clínica. O desempenho do nomograma foi avaliado por medidas globais, de calibração e de discriminação..

Resultados: O nomograma mostrou a discriminação moderada e boa calibração (AUC 0,79; escore Brier 0,064, ângulo de discriminação 0,13; Hosmer-Lemeshow p = 0,27)

Conclusões: as probabilidades preditas mostraram concordância global com a freqüência observada de pneumonia após cirurgia cardíaca. O nomograma forneceu uma predição satisfatória da probabilidade de pneumonia. Sua aplicabilidade para o uso clínico pode contribuir para que pacientes e cirurgiões tenham uma melhor avaliação sobre o risco de pneumonia pós cirurgia cardíaca, antes da realização do procedimento.

Introdução

Pneumonia nosocomial é uma complicação freqüente em medicina intensiva e está associada com aumento do custo da atenção médica e mortalidade [1]. Pacientes de cirurgia cardíaca constituem uma população de alto risco para a pneumonia, devido aos danos mecânicos para os pulmões, causados pelo procedimento[2]. Regras de predição combinam dados clínicos e laboratorias para prever diversos desfechos como óbito e infecções. Prever a gravidade das complicações com o uso de fatores de risco pré-operatório pode ser importante para o paciente e cirurgião, particularmente na definição de medidas preventivas.

A maioria das regras de predição foi derivada, mas não validada, ou foi limitada a dividir as amostras, levando a possibilidade de otimismo do desempenho e modelos superajustados aos dados [3].

Nomogramas têm sido desenvolvidos para melhorar a aplicabilidade dos modelos preditivos. Eles fornecem uma representação gráfica da força de preditores individuais, e permitem aos clínicos calcular uma pontuação para pacientes individuais, refletindo o seu risco pessoal. O objetivo do presente estudo foi projetar um nomograma que possa predizer a ocorrência de pneumonia pós-cirurgia cardíaca com fatores de risco pré-operatórios convencionais e amplamente disponíveis.

Pacientes e Métodos

Modelo de Derivação

O modelo de derivação foi desenvolvido entre 2000 e 2002 em um pequeno hospital privado. Os dados (amostra de derivação) foram coletados de 527 pacientes consecutivos que foram submetidos a grandes cirurgias cardíacas (revascularização miocárdica, cirurgia valvar, ou cirurgia corretiva de cardiopatia congênita em adultos). Os pacientes foram excluídos quando apresentaram pneumonia em até 15 dias antes da cirurgia ou evoluíram para óbito nas primeiras 48 h após a cirurgia. Parâmetros dos pacientes foram prospectivamente registrados e armazenados em um banco de dados computadorizado.

Este protocolo foi aprovado pelo Comitê de Ética da Instituição sob o número de investigação # 0111/17.7. 06. Pneumonia foi diagnosticada durante a vigilância ativa por um especialista em doenças infecciosas (MS), de acordo com os critérios publicados pelo *Center for Disease Control and Prevention*[4]. A confirmação microbiológica do diagnóstico não era necessária. Outros detalhes metodológicos foram descritos anteriormente [5]

Variáveis de exposição

Somente variáveis pré-operatórias foram incluídas nos modelos preditivos. O conjunto das seguintes variáveis foi avaliado: **demográficas** (sexo, idade, índice de massa corporal,

peso, altura); **comorbidades**: (diabetes, doença vascular periférica, doença pulmonar obstrutiva crônica, tabagismo atual, creatinina sérica, necessidade de diálise, função ventricular, classe da *New York Heart Association*, parâmetros funcionais, regurgitação mitral, angina instável, tamanho do átrio esquerdo, hipertensão pulmonar, infarto agudo do miocárdio e arritmia); **fatores cirúrgicos** (duração da estadia no hospital antes da cirurgia, cirurgia prévia e classificação como cirurgia em eletiva, urgente ou emergencial).

Análise Estatística

A regressão logística múltipla foi utilizada para avaliar associações entre potenciais variáveis preditoras e a ocorrência de pneumonia. As variáveis com valor de $p = 0,2$ para a predição de pneumonia foram inseridas no modelo multivariado. Um procedimento de eliminação manual foi então realizado, e um valor de p de 0,05 foi o critério para permanência no modelo. Um nomograma foi desenvolvido para visualizar o prognóstico dos diferentes fatores prognósticos em um único diagrama. O nomograma foi então usado para calcular a incidência esperada de pneumonia com base em um determinado perfil de risco do paciente. Três categorias de medidas de desempenho foram utilizadas: discriminação, calibração e medidas globais. A discriminação foi definida como a capacidade de distinguir corretamente entre os dois fatores que foram e não foram associados com o resultado. A calibração foi definida como uma medida de quão perto as probabilidades preditas concordaram com os resultados reais. As medidas globais incluíram tanto medidas de discriminação quanto de calibração. Gráficos de calibração foram usados para comparar a adequação do ajuste e a validade das probabilidades preditas.

A validade interna do modelo foi avaliada por meio de técnicas de *bootstrap*[6]. *Bootstrapping* é um método de reamostragem que permite inferências a serem feitas sobre a população original de desenho substituindo amostras da população original. A diferença entre o desempenho da amostra *bootstrap* e o desempenho da amostra de derivação foi considerada uma estimativa do otimismo no resultado do desempenho da derivação, ou seja o quanto os resultados são "inflados" em relação ao valor real. Essa diferença foi calculada para cada amostra *bootstrap*. A média das diferenças foi utilizada para obter uma estimativa estável do otimismo. O otimismo foi subtraído do desempenho aparente na derivação para estimar o desempenho validado internamente. Para estimar o intervalo de

confiança de 95% (CI), usamos as distribuições empíricas das 200 amostras *bootstrap*[7]. Os dados foram analisados com Stata (software de estatística, versão 9.0 para *Windows, Stata Corp, College Station, Texas*) e R (versão 2.7.1 *Copyright C 2008 The R Foundation for Statistical Computing*) para a análise exploratória. Desempenho do modelo foi avaliado através da adaptação de um código de R originalmente criado por *Steyerberg*[7], e o nomograma foi criado a partir da Biblioteca de projetos com R-software[8]. O script de R é mostrado no Anexo 1.

Resultados

Pneumonia ocorreu em 7,5% dos pacientes na amostra de derivação. Foram excluídos 1,5% dos pacientes considerados para o estudo devido a dados incompletos. A população era dois terços pertencentes ao sexo masculino, cerca de metade dos pacientes tinham angina instável, e cerca de 10% necessitaram de cirurgias de emergência (Tabela 1).

As características da população refletem um pequeno hospital privado com perfil de um serviço de emergência aberto, pacientes idosos e grande prevalência de angina instável (Tabela 1).

Tabela 1 - Distribuição das variáveis preditoras na amostra de derivação (n=527)

Variable	
Idade: media (anos)	64
Sexo feminino %	33
Tempo de permanencia pré-operatório em dias (intervalo)	3,68 (0-305)
Regurgitaçao mitral %	9,11
Angina instável %	48,2
Revascularizaçao prévia %	7,02
Creatinina média mg/dl	0,9
Diabetes mellitus %	23,91
Dialise %	2,80
Doença Vascular Periférica %	30,74
Doença pulmonar obstrutiva cronica %	7,40
Tamanho atrio esquerdo: media mm	3,70
Fumo %	11,20
Hipertensão pulmonar (%)	5,31
Infarto agudo do miocárdio (%)	
<3 meses antes da cirurgia	15,87
Índice de Massa Corporal (kg/m^2): media	26,21
Classe IV *New York Health Association* (%)	12,79

Peso, media Kg	72,90
Ritmos não sinusal (%)	8,16
Cirurgia de emergência (%)	10,06
Disfunção Ventricular Moderada/Grave (%)	15,62

Método

Quatro variáveis de fácil obtenção foram selecionadas para o modelo (Tabela 2): A doença pulmonar obstrutiva crônica (DPOC), definida como uma história de DPOC ou DPOC diagnosticada com base nos exames complementares, ou uso de corticóides e/ou broncodilatadores; idade (anos); a necessidade de cirurgia de emergência e a função ventricular esquerda, definida como: normal, disfunção leve ou moderada / grave com base nos achados ecocardiográficos ou observações durante o cateterismo. O tipo de cirurgia cardíaca foi considerado eletivo quando o procedimento pode ser adiado sem aumentar o risco de lesão cardíaca, de urgência quando o procedimento foi necessário durante a internação para minimizar as chances de deterioração clínica posterior, ou de emergência quando o procedimento foi necessário imediatamente (causado por: disfunção isquêmica, isquemia progressiva; infarto agudo que ocorreu até 24h antes da cirurgia, edema pulmonar agudo, disfunção mecânica, choque com ou sem suporte circulatório). O tipo de cirurgia foi transformado em uma variável dicotômica, sendo cirurgia de emergência o preditor mais forte dentre as variáveis.

Tabela 2 - Modelo de regressão logística

Variavel	**Categoria**	**Razão de chances**	**IC 95%**	***p***
DPOC	sim	4,29	1,73 – 10,61	0,02
	não	1,00		
Cirurgia de Emergência	sim	5,58	2,56-12,12	<0,01
	não	1,00		
Idade	Ordinal	1,04	1,0 - 1,08	0,02
Disfunçao ventricular moderada/grave	sim	2,68	1,27- 5,66	0,01
	não	1,00		

Um nomograma foi criado a partir de uma série de gráficos empilhados que exibem os intervalos dos riscos preditos (Figura 1). A primeira linha (PONTOS) forneceu uma atribuição de pontos para cada variável, conforme a força do preditor defeinida pela regressão logística. Nas segunda e quinta linhas são representadas as variáveis incluídas no modelo. Para um paciente individual, cada valor de variável foi convertido em um valor de ponto, traçando-se uma linha vertical em relação ao valor apropriado da variável para a linha de pontos. Em seguida, os pontos atribuídos para as quatro variáveis foram somados, e o valor total foi localizado na chamada escala total de pontos. A partir dali, uma linha vertical foi traçada com o valor correspondente ao da linha de fundo (Risco de Pneumonia). Por exemplo, um paciente de 60 anos, com a cirurgia de emergência, sem disfunção ventricular e com DPOC tiveram uma pontuação total de 50 +53 +0 +44 = 147 pontos no total e, a probabilidade estimada foi de aproximadamente 0,42.

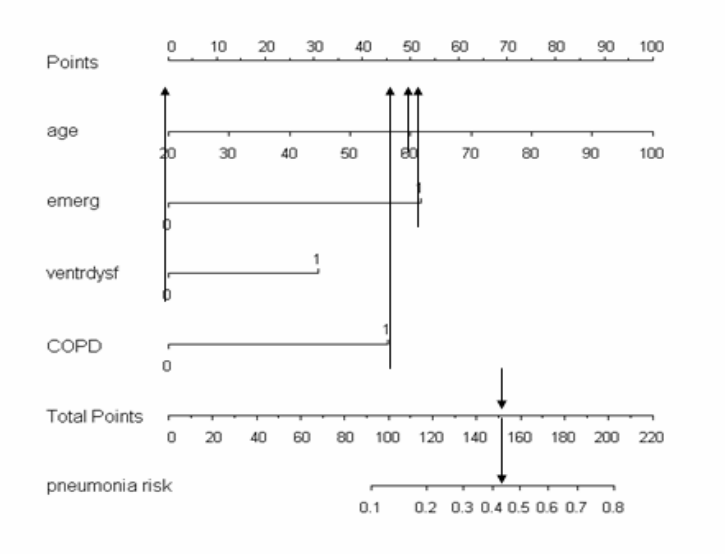

Figura 1 - Nomograma com uma estimativa de um exemplo particular

Medidas de desempenho

O valor observado do índice C, correspondente à Área sob a curva ROC, para um valor ideal de 1, o valor observado foi de 0, 798 para o conjunto de derivação e 0,79 para o valor de validação interna (Tabela 3), indicando boa concordância entre o predito e o

observado. A capacidade de discriminação foi pequena quando avaliado pelo ângulo de discriminação (ideal 1). O ângulo de calibração ideal era 1, e foi de 0,92 considerado adequado. A adequação das medidas de ajuste (testes de *Hosmer-Lemeshow e Goeman le Cessie*) obteve um p-valor ideal acima de 0,05. O resultado do R^2 mostrou que apenas uma pequena quantidade da variação nos dados foi explicada pelo modelo, e o escore de *Brier*, que quantifica o erro de previsão, foi baixo, tanto no conjunto de derivação e do conjunto de validação interna, refletindo um bom desempenho global.

Análise gráfica

A calibração e a discriminação também foram avaliadas graficamente. O gráfico de discriminação mostrou boa capacidade de discriminação (Figura 2) e a curva de calibração mostrou que a calibração se deteriorou para probabilidades acima de 20%, com uma probabilidade predita maior do que a probabilidade observada nesse grupo (Figura 3). Portanto nos cenários com prevalência de pneumonia acima de 20% o modelo teria ótima calibração.

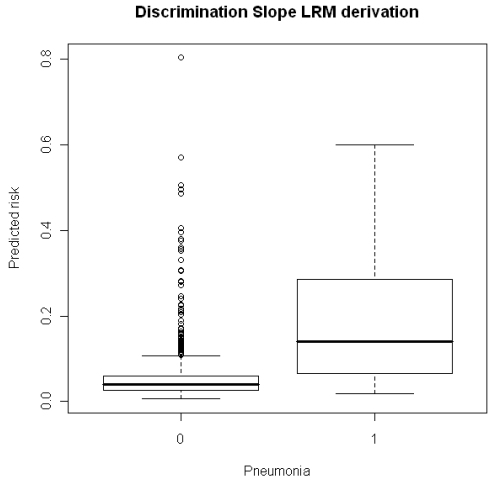

Figura 2 - Gráfico de discriminação

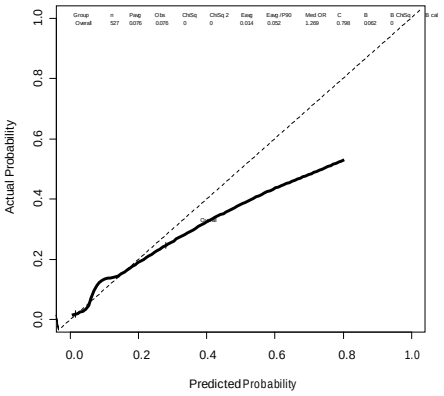

Figura 3 - Gráfico de calibração

Seção 5.01 Medida	**Derivação**	**Validação Interna**	***Seção 5.02 Valor Ideal***
Tabela 3 - Medidas de desempenho modelo de predição de pneumonia			
R^2	20,9%	18%	100%
Brier	0,062	0,064	0
indice C	0,798	0,79	1
IC 95%	[0,72 - 0,88]	[0,71 - 0,86]	-
Angulo de Discriminação	0,13	0,13	100
IC 95%	0,08 - 0,18	0,07 - 0,18	-
Calibração em larga escala	0	0	0
Angulo de Calibração	1	0,92	1
Hosmer- Lemeshow	0,81	-	>0,05
Goeman le Cessie	p=0,27	-	>0,05

Discussão

A discriminação se refere à capacidade de distinguir pacientes de alto risco dos pacientes de baixo risco e pode ser avaliada com o índice C (área sob a curva ROC) e com o ângulo de discriminação. Em nosso modelo de predição de pneumonia, o índice C e o ângulo de discriminação mostraram pequenas diferenças entre os dados de derivação e os dados de validação interna. Calibração refere-se ao fato das probabilidades preditas concordarem com as probabilidades observadas. Porque os riscos não podem ser diretamente observados, somente caso o paciente tivesse pneumonia ou não, a calibração só poderia ser medida indiretamente, seja com grupos ou com a hipótese de que seria testado se o modelo está bem calibrado ("encaixa") para os dados [9]. É representada graficamente por uma linha (Figura 3), que pode ser descrita com uma inclinação e um intercepto (o ângulo de calibração e a calibração em larga escala, respectivamente). No conjunto de derivação, o valor do ângulo de calibração foi determinado pelo coeficiente de regressão. Em nosso modelo de predição de pneumonia, os dados de validação interna realizam uma predição com otimismo, indicando que as previsões eram extremas em novos pacientes, demasiado baixas em pacientes de baixo risco ou muito elevadas em pacientes de alto risco[10]

As medidas globais se referem tanto à calibração quanto à discriminação, e incluem o R^2 e escore de *Brier*. O modelo de predição de pneumonia mostrou bom desempenho global (*Brier* com pontuação baixa), mas explica apenas uma parcela pequena da variação (R^2= 10%).

Os modelos de predição pré-operatória têm três usos principais: promover uma eficiente organização intra-hospitalar, informar o paciente, e facilitar o ajuste de risco. A organização intra-hospitalar pode ser melhorada com o conhecimento pré-operatório da probabilidade de complicações. Por exemplo, a lista de espera de cirurgia poderia ser priorizada com base na probabilidade de complicações, os recursos de terapia intensiva poderiam ser devidamente alocados, e os pacientes em grupos de alto risco poderiam receber estratégias preventivas específicas ou mais intensas. Esta eficiência organizacional poderia melhorar a relação custo-benefício da assistência hospitalar. A informação ao paciente também pode ser melhorada com o conhecimento pré-operatório da probabilidade de complicações. Paciente e familiares podem tomar melhores decisões com as previsões realistas das probabilidades de complicações pós-operatórias, especialmente em cirurgias de alta complexidade, como cirurgia cardíaca. A ocorrência de pneumonia no pós-operatório é sempre associada a um impacto psicológico e econômico. Potencial alto custo das intervenções médicas inclui ventilação mecânica prolongada, traqueostomia, e aumento do tempo de permanência em unidade de cuidados intensivos.

A utilização de previsão pré-operatória como uma ferramenta de ajuste de risco é limitada ao controle de infecção. As regras de predição do sistema NNIS (*The National Nosocomial Infections Surveillance (System)*[4] são as mais utilizadas no controle de infecção, apresentando uma série de limitações nas previsões de infecção da ferida cirúrgica, especialmente em cirurgia cardíaca. Não há regra de predição que tenha sido amplamente adotada para pneumonia. As regras de mortalidade do *Acute Physiology Age Chronic Health Evaluatio*n (APACHE) têm sido utilizadas com mais freqüência em unidades de terapia intensiva brasileiras. A variação mensal na gravidade dos casos torna difícil fazer uma avaliação da qualidade das taxas de pneumonia. O advento de um conjunto de regras de predição de sucesso para pneumonia iria facilitar as comparações entre o esperado e o observado, facilitar a detecção de taxas anormais, e o desenho de intervenções de prevenção.

Um nomograma pode ser usado para visualizar graficamente a força prognóstica dos fatores de risco e calcular a probabilidade de pneumonia esperada, com base em um perfil individual de simples características pré-operatórias. Quando externamente validada, o nomograma pode ser usado para predição de morbidade pós-cirurgia cardíaca. O nomograma não requer recursos computacionais para uso clínico, já que envolve simplesmente o desenho de linhas e soma dos pontos dos fatores individuais para avaliar o risco em cirurgia cardíaca. O modelo nomograma desenvolvido neste estudo utiliza apenas preditores pré-operatórios e mostrou discriminação limitada, mas uma boa calibração. *Briganti* [11] mostrou

que os clínicos estão mais familiarizados com nomogramas do que com a consulta de tabelas ou árvores, e os nomogramas apresentaram taxas mais elevadas na utilidade clínica.
Nomogramas são amplamente aplicados à predição de recorrência do câncer, mas também têm sido desenvolvida para predizer a probabilidade de desenvolver complicações maiores após esofagectomia, reconstrução da mama, metástase no câncer de mama e cirurgia oftalmológica [12-15], além disso, o nomograma normalmente supera o julgamento clínico[16]. O modelo estudado inclui quatro fatores pré-operatórios. Esses fatores limitam as intervenções em potencial, porque eles não podem ser removidos, no entanto, algumas estratégias de prevenção são possíveis. Por exemplo, *Enterobacter sp*, considerado um importante patógeno isolado em pneumonia após a cirurgia cardíaca, sendo resistente às cefalosporinas de primeira geração. Selecionar um antibiótico profilático de maior espectro para pacientes de alto risco pode limitar o impacto ecológico, por exemplo, a profilaxia com quinolonas.

Outro importante objetivo deste estudo foi descrever um novo método para a avaliação do desempenho do modelo. As publicações nesta área são limitadas a revistas epidemiológicas, e a maioria dos médicos limita as avaliações de modelos ao valor do índice C. O índice C[17] não é afetado pela prevalência da doença, mas depende da variedade de casos e da gravidade da doença, assim, este teste pode ser sensível entre pacientes com doença grave e pouco efetivo fora deste cenário. Foi elaborado um roteiro em R para implementar estes métodos que estão disponíveis gratuitamente para acesso fácil pelos médicos.

Este estudo tem algumas limitações. Em primeiro lugar, de acordo com *Guyatt et al.*[18], todas as regras de predição devem ser validadas externamente. Neste estudo, utilizou-se de validação interna, que só considera pura amostragem, variabilidade com técnicas *bootstrap;* não considera alterações na população de pacientes *(''case-mix)[3]*. Embora a calibração do modelo fosse boa, este tipo de validação não garante a sua generalização, no qual o desempenho do modelo é analisado em outros ambientes com diferentes tipos de processos e procedimentos. O modelo tem apenas uma capacidade limitada de discriminação, que pode ser atribuída ao fenômeno multifatorial e complexo inerente às infecções hospitalares, que incluem, além de fatores do paciente, incidentes cirúrgicos e atitudes dos profissionais que não são possíveis de prever no pré-operatório. Nosso objetivo em definir medidas preventivas para pacientes de alto risco também limitou o modelo em utilizar apenas fatores pré-operatórios. Outra possível limitação foi a utilização de variáveis que fossem facilmente obtidas para facilitar a aplicabilidade clínica do modelo. Finalmente, pode ter sido

introduzido um viés de classificação, porque um diagnóstico de pneumonia no pós-operatório de cirurgia cardíaca representa um desafio clínico.

Em conclusão, a cirurgia cardíaca é acompanhada de alto risco de infecção no pós-operatório. O modelo aqui descrito mostrou um nomograma com discriminação moderada, mas boa calibração tanto no desenvolvimento quanto na validação interna. A simplicidade do nomograma facilita sua aplicação na prática clínica para avaliação de risco em cirurgia cardíaca. O nomograma pode facilitar a seleção de uma estratégia de cirurgia ideal e adequada, além de otimizar a escolha pré-operatória de medidas preventivas. Quando externamente validado, o nomograma pode desempenhar um papel no ajuste de risco de morbidade pós-cirurgia cardíaca.

(i) Referências

[1] Kollef MH, Sharpless L, Vlasnik J, Pasque C, Murphy D, Fraser VJ. The impact of nosocomial infections on patient outcomes following cardiac surgery. Chest. 1997 Sep;112(3):666-75.

[2] Beloborodova NV, Nonikov VE, Bachinskaia EN. [Diagnostic and therapeutic aspects of pneumonias after cardiosurgical operations associated with artificial ventilation]. Anesteziologiia i reanimatologiia. 2006 May-Jun(3):83-7.

[3] Bleeker SE, Moll HA, Steyerberg EW, Donders AR, Derksen-Lubsen G, Grobbee DE, et al. External validation is necessary in prediction research: a clinical example. Journal of clinical epidemiology. 2003 Sep;56(9):826-32.

[4] Horan TC, Culver DH, Gaynes RP, Jarvis WR, Edwards JR, Reid CR. Nosocomial infections in surgical patients in the United States, January 1986-June 1992. National Nosocomial Infections Surveillance (NNIS) System. Infect Control Hosp Epidemiol. 1993 Feb;14(2):73-80.

[5] Santos M, Braga JU, Gomes RV, Werneck GL. Predictive factors for pneumonia onset after cardiac surgery in Rio de Janeiro, Brazil. Infect Control Hosp Epidemiol. 2007 Apr;28(4):382-8.

[6] Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. Journal of clinical epidemiology. 2001 Aug;54(8):774-81.

[7] Steyerberg EW. Clinical prediction models : a practical approach to development, validation, and updating. New York ; London: Springer 2009.

[8] Harrell FE, Jr., Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: advantages, problems, and suggested solutions. Cancer treatment reports. 1985 Oct;69(10):1071-77.

[9] Peek N, Arts DG, Bosman RJ, van der Voort PH, de Keizer NF. External validation of prognostic models for critically ill patients required substantial sample sizes. Journal of clinical epidemiology. 2007 May;60(5):491-501.

[10] Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. Journal of clinical epidemiology. 2008 Jan;61(1):76-86.
[11] Briganti A, Capitanio U, Suardi N, Gallina A, Rigatti P, Montorsi F. Re: Michael W. Kattan. Classification and Regression Trees Versus Nomograms: A Bone Scan Positivity Example. Eur Urol. In press. doi:10.1016/j.eururo.2010.01.005. European urology. Jan 20.
[12] Chun FK, Karakiewicz PI, Briganti A, Walz J, Kattan MW, Huland H, et al. A critical appraisal of logistic regression-based nomograms, artificial neural networks, classification and regression-tree models, look-up tables and risk-group stratification models for prostate cancer. BJU international. 2007 Apr;99(4):794-800.
[13] Lagarde SM, Reitsma JB, Maris AK, van Berge Henegouwen MI, Busch OR, Obertop H, et al. Preoperative prediction of the occurrence and severity of complications after esophagectomy for cancer with use of a nomogram. The Annals of thoracic surgery. 2008 Jun;85(6):1938-45.
[14] Van Zee KJ, Manasseh DM, Bevilacqua JL, Boolbol SK, Fey JV, Tan LK, et al. A nomogram for predicting the likelihood of additional nodal metastases in breast cancer patients with a positive sentinel node biopsy. Annals of surgical oncology. 2003 Dec;10(10):1140-51.
[15] Subbaram MV, MacRae SM. Customized LASIK treatment for myopia based on preoperative manifest refraction and higher order aberrometry: the Rochester nomogram. J Refract Surg. 2007 May;23(5):435-41.
[16] Ross PL, Gerigk C, Gonen M, Yossepowitch O, Cagiannos I, Sogani PC, et al. Comparisons of nomograms and urologists' predictions in prostate cancer. Seminars in urologic oncology. 2002 May;20(2):82-8.
[17] Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007 Feb 20;115(7):928-35.
[18] Guyatt G. Users' guides to the medical literature : essentials of evidence-based clinical practice. 2nd ed. New York: McGraw-Hill Medical 2008.

Anexo A Construção do nomograma

```
require(Design)
require(foreign)
require(ROCR)
require(Hmisc)
require(epiR)
setwd("C:banco de dados")
#LRM
# A-Derivation Model
#READING FILE
DEVPRO<-read.dta("banco_derivacaoPRO_21nov.dta")
#building model
DEVPRO.lrm <- lrm(vap ~ age+COPD+emerg+ventrdysf, data=DEVPRO, x=T,y=T,linear.predictors=T)
#2-extracting linear predictors
lp.lrm <- DEVPRO.lrm$linear.predictors
#3-diagnosis
DEVPRO.lrm$y <- as.numeric(DEVPRO.lrm$y)
val.prob(logit=DEVPRO.lrm$linear.predictors, y=DEVPRO.lrm$y) #model diagnosis
title(paste("calibration graph- LRM derivation"))
#calibration diagnosis
LR<-val.prob(logit=DEVPRO.lrm$linear.predictor, group=T, y=DEVPRO.lrm$y)
plot(LR)
#6-CI C index
cstat.dev.lrm <- rcorr.cens(DEVPRO.lrm$linear.predictors, DEVPRO.lrm$y)
cat(cstat.dev.lrm[1], "[", cstat.dev.lrm[1]-1.96/2*cstat.dev.lrm[3], " - ",
cstat.dev.lrm[1]+1.96/2*cstat.dev.lrm[3],"]")
#7-discrimination slope and boxplot
mean(plogis(lp.lrm[DEVPRO.lrm$y==1])) - mean (plogis(lp.lrm[DEVPRO.lrm$y==0]))# calculanting slope
boxplot(plogis(lp.lrm)~ DEVPRO.lrm$y,ylab="Risco predito", xlab="Pneumonia")
title(paste("Discrimination Slope LRM derivation"))
#8-calculating sd #erro
# Bootstrap discrimination slope
nrowB   <- nrow(DEVPRO)  # nrow from development set
B <- 200          # 200 bootstraps
matB <- matrix(NA,nrow=B,ncol=3) # Matrix for results
dimnames(matB) <- list(c(1:B), Cs(Slopeapp, Slopetest, optimism ))
# Start loop
for (i in 1:B) {
if (i%%10==0) cat("Start Bootstrap sample nr", i, "\n")
Brows <- sample(nrowB,replace=T)
```

```
# Bsample is bootstrap sample from development set
Bsample<- DEVPRO[Brows,]
devfull.lrm <- lrm(vap ~ age+COPD+emerg+ventrdysf, data=Bsample,linear.predictors=T, x=T, y=T)
matB[i,1] <- mean(plogis(devfull.lrm$linear.predictors[devfull.lrm$y==1])) -
        mean(plogis(devfull.lrm$linear.predictors[devfull.lrm$y==0]))
lp.b.lrm  <- DEVPRO.lrm$x %*% devfull.lrm$coef[2:length(devfull.lrm$coef)] + devfull.lrm$coef[1] # lp with
coefs from bootstrap
matB[i,2]  <-  mean(plogis(lp.b.lrm[DEVPRO.lrm$y==1])) -  mean(plogis(lp.b.lrm[DEVPRO.lrm$y==0]))    #
Testing on original sample
matB[i,2]
} # End for loop
cat("\n\n\n\n")
matB[,3] <- matB[,1] - matB[,2] # optimism
apply(matB,2,mean)
apply(matB,2,sd)
k1 <- apply(matB,2,mean)
k2 <- apply(matB,2,sd)
#5-CI discrimination slope
k3  <-mean(plogis(lp.lrm[DEVPRO.lrm$y==1])) -  mean  (plogis(lp.lrm[DEVPRO.lrm$y==0]))# calculando a
slope
cat(k3, "[", k3-1.96*k2[2], " - ", k3+1.96*k2[2],"]")
#8-H-L
source("HLtest.r")
DEVPRO.lrm$y<- as.numeric(DEVPRO.lrm$y)
hl.ext2(p= plogis(DEVPRO.lrm$linear.predictor),y=DEVPRO.lrm$y,g=10,df=8)
#9-Goeman le Cessie
source("multinomialcode.R")#download script
lform <- vap ~ age+COPD+emerg+ventrdysf #including the model
fit <- mlogit(lform, DEVPRO) #multinomial
U <- smoothU(DEVPRO, c("age", "COPD","emerg","ventrdysf"), 0.25,"k")#smoothing of residuals
testfit(fit, U %*% t(U)) #output of test, first value = p value (ideal>0.05)
#Internal validation
DEVPRO.lrm <- lrm(vap ~ age+COPD+emerg+ventrdysf, data=DEVPRO, x=T,y=T,linear.predictors=T)
val.int <- validate(DEVPRO.lrm, B=200) #internal validation
val.int
castat.val <- (val.int[1,5]*0.5)+0.5
cat(castat.val, "[", castat.val-1.96/2*cstat.dev.lrm[3], " - ", castat.val+1.96/2*cstat.dev.lrm[3],"]")
#creating  nomogram
nomo<- read.dta("banco_derivacaoPRO_21nov.dta")
attach(nomo)
```

```
ddist<-datadist(age,ventrdysf,emerg,COPD)
options(datadist='ddist')
f<-lrm(vapn~age+emerg+ventrdysf+COPD)
nomogram(f,fun=plogis,funlabel="pneumonia risk", lp=F, xfrac=.25)
```

5 ARTIGO III - VALIDAÇÃO EXTERNA DE MODELOS PARA PREDIÇÃO DE PNEUMONIA PÓS CIRURGIA CARDÍACA

Resumo

Infecções graves são complicações importantes pós-cirurgias cardíacas. Predizer a ocorrência destas infecções é útil para sua prevenção.
Objetivo: Validar e comparar modelos de predição estimando o risco dos pacientes desesnvolverem pneumonia pós-cirurgia cardíaca.
Desenho: Estudo prospectivo para validar modelos de predição de pneumonia pós- cirurgias cardíacas.
Pacientes: Adultos (n=333) submetidos a cirurgias cardíacas entre Outubro de 2006 a Maio de 2007.
Métodos: Modelos de regressão logística (LRM) e árvore de classificação e regressão (CART). foram validados com dados externos .
Resultados: Pneumonia ocorreu em 7,5% dos pacientes na amostra de validação. LRM apresentou melhor desempenho com moderada discriminação (R2 7,1%, Brier=0.06, AUC=0, 694) e calibração adequada (Hosmer-Lemeshow P=0,08).
Conclusões: Foi validado um modelo capaz de identificar pacientes de alto risco para pneumonia, submetidos a cirurgias cardíacas. CART apresentou um bom desempenho na derivação e maiores perdas do que LRM, quanto à discriminação e calibração, na amostra de validação.

Introdução

Infecções comuns pós-cirurgia cardíaca aumentando o tempo de permanência na unidade de terapia intensiva (UTI), a letalidade e os custos. Aproximadamente, 20% dos pacientes têm infecções pós-cirurgia cardíaca [1]. A pneumonia é uma das mais graves complicações pós-operatórias, aumentando a chance de morte em 14 vezes[2].

A maior freqüência de pneumonia pós-cirurgia cardíaca é explicada pelas condições cirúrgicas como: circulação extracorpórea; imunossupressão; presença de drenos torácicos; dor levando a hipoventilação; trauma pulmonar e ausência de ventilação pulmonar durante a cirurgia. A prevenção da pneumonia é uma parte importante desse cenário. A seleção de grupos de pacientes de alto risco pode ajudar em muitas estratégias, incluindo a intervenção seletiva em grupos de maior risco, o ajuste de taxas, aperfeiçoamento da programação cirúrgica, e informação sobre o risco cirúrgico.

A maioria dos trabalhos que discute as regras de predição está limitada ao desenvolvimento das regras ou derivação. Um pequeno número de estudos aborda a validação de regras, e raramente se aborda a incorporação e impacto sobre o comportamento do médico ou desfecho no paciente [3]. A falta de validação externa pode representar um problema, porque uma boa calibração e discriminação na amostra de derivação não garantem uma predição adequada em novos pacientes. A maioria das regras de predição mostra redução na acurácia quando testada em novos pacientes[4].

O objetivo do presente estudo foi comparar o desempenho de dois modelos de predição, a árvore de classificação e regressão (CART) e um modelo de regressão logística (LRM) sobre uma amostra de validação externa.

Métodos

A LRM e o CART foram construídos em um estudo anterior[5] com 527 pacientes adultos submetidos à cirurgia cardíaca entre junho de 2000 e agosto de 2002, em um pequeno hospital privado. Os pacientes foram prospectivamente analisados para identificar fatores prognósticos para a pneumonia. A pneumonia ocorreu em 7,5% dos pacientes na amostra derivação.

Uma amostra de validação foi recrutada em um hospital público, o Instituto Nacional de Cardiologia (INC), registrando-se 333 pacientes consecutivos, adultos, submetidos a qualquer cirurgia cardíaca de grande porte, entre outubro de 2006 e maio de 2007. O protocolo de pesquisa foi aprovado pelo comitê de ética (número de investigação 0111/17.7. 06)

Seleção de Pacientes

Todos os pacientes submetidos a cirurgias cardíacas (revascularização miocárdica, cirurgia valvar de adultos e cardiopatias congênitas do adulto), que sobreviveram por pelo menos 48 horas foram incluídos. Os pacientes submetidos a procedimentos cirúrgicos menores, como drenagem e biópsia pericárdica, e aqueles que apresentaram pneumonia até 15

dias antes da cirurgia, foram excluídos. Os prontuários não estavam disponíveis para 0,6% dos pacientes considerados para inclusão no estudo.

Critérios de diagnóstico e definição de fatores preditivos

Pacientes definidos como tendo pneumonia pós-operatória apresentaram critérios de definição de pneumonia hospitalar padronizados pelo *CDC*[6], tendo sido avaliados por um especialista em doenças infecciosas (MS) "cego" em presença de fatores de predição de pneumonia. A confirmação microbiológica do diagnóstico não era necessária. As variáveis demográficas foram: sexo (masculino ou feminino), idade (anos), peso (kg medido ou relatado). Doenças pré-existentes incluíram: doença pulmonar obstrutiva crônica (DPOC), definida como uma história de DPOC, DPOC diagnosticada com base nos resultados de exames complementares, ou o uso de corticóides (para doença pulmonar) e/ou broncodilatadores; função ventricular, e que foi dicotomizada como (a) normal ou disfunção leve e (b) disfunção moderada ou grave, com base em resultados do ecocardiograma ou cateterismo.

A classificação da cirurgia cardíaca (características cirúrgicas) foi definida como eletiva (o procedimento poderia ser adiado, sem aumentar o risco de dano cardíaco), urgência (o procedimento deveria ser realizado durante a internação para minimizar a probabilidade de deterioração clínica posterior, na ausência de indicação eletiva ou de situação de emergência), ou de emergência (o procedimento seria necessário de imediato devido a uma disfunção isquêmica, definida como isquemia progressiva; infarto agudo nas 24 horas anteriores à cirurgia; edema pulmonar agudo, necessitando de intubação, ou disfunção mecânica, definida como choque com ou sem suporte circulatório). O tipo de cirurgia foi transformado em uma variável dicotômica, emergência *versus* não-emergência (eletiva+urgência).

Análise Estatística

Os dados foram analisados utilizando-se dois pacotes de *software* para análise exploratória: *Stata statistical software* (versão 9.0 para *Windows, Corp Stata, College Station, Texas*) e R (versão 2.7.1, a Fundação para R *Statistical Computing*). Para o desempenho do modelo foi feita a adaptação de um código de R originalmente criado por *Steyerberg*[7].

Os métodos estatísticos tradicionais são de difícil utilização ou de utilidade limitada na resolução de problemas de classificação. A regressão logística é o método mais comum utilizado na análise de dados médicos [8], mas apresenta algumas dificuldades como preditor de distribuições variáveis, interações complexas e padrões. Ao longo dos últimos anos, tem havido um crescente interesse no uso da análise por CART, uma técnica de construção de árvore que é diferente de métodos tradicionais de análise de dados. Devido a esta diferença, a CART foi aceita de forma relativamente lenta. *Lewis*[9] argumentou que a CART é bastante eficaz para a criação de regras de decisão clínica e executa tão bem, ou melhor, do que as regras desenvolvidas, utilizando métodos mais tradicionais, além de ser capaz de descobrir as interações complexas entre os indicadores que podem ser difíceis ou impossíveis de descobrir, utilizando técnicas tradicionais multivariadas. Para desfechos dicotômicos a análise por CART é usada para construir um sistema de classificação binária baseada na presença ou ausência do desfecho em questão através de partições sucessivas, dividindo os dados em subgrupos mais homogêneos, com cada divisão, ou "nó". Em cada nó, o algoritmo seleciona a variável com maior capacidade de discriminar entre dois resultados, neste caso, a presença ou ausência de pneumonia. A árvore de classificação tem uma estrutura hierárquica em que a primeira divisão corresponde à variável com maior poder discriminatório.

As medidas de desempenho foram agrupadas em três categorias (Tabela 1) discriminação, calibração e medidas globais. A discriminação se refere à capacidade de separar corretamente dois grupos: com e sem o desfecho. A calibração representa o quão estreitamente, probabilidades preditas concordam com os resultados reais. Medidas Globais incluem ambas as habilidades anteriores.

Tabela 1- Medidas de desempenho

Medida	Cálculo	Prós	Contras
Medidas Globais			
R^2	Logaritmo das predições: compara com o desfecho real Ideal: 100%	Fácil interpretação, variação explicada devido às variáveis preditoras	Penaliza muito os valores extremos com desfechos discordantes
Escore de Brier	Diferença quadrática entre desfechos reais e preditos (Y-Ŷ)2 Ideal: zero	Pode ser usado para qualquer tipo de desfecho (binário, contínuo, censurado)	Afetado pela baixa prevalência
Medidas de Discriminação			
Estatística de Concordância (índice C)	Estatística ranqueada para predições, Probabilidade de um indivíduo escolhido ao acaso com o desfecho tenha uma predição maior do que o sem o desfecho, Corresponde a AUC Ideal: 1	Uso clínico freqüente. Interpretação visual com a curva ROC. Não é afetada pela prevalência.	Pouco sensível, depende das diferenças de gravidade dos pacientes da amostra.
Ângulo de discriminação	Diferença absoluta entre a predição média dos indivíduos com e sem desfecho Ideal: 100	Interpretação visual com o gráfico de discriminação	Afetado pela baixa prevalência
Medidas de Calibração			
Ângulo de calibração	Angulo da regressão em um preditor linear Ideal: 1	Fator de "encolhimento" para ajuste do modelo reflete o superajuste e diferenças no efeito dos preditores	Interpretação difícil. Útil apenas na validação externa
Teste de *Hosmer-Lemeshow* (HL)	Compara desfechos preditos e observados em grupos de pacientes (teste chi2) Ideal p>0,05	Conceito simples	Pequeno poder em amostras reduzidas Falha em detector superajuste dos predictores
Teste de *Goeman le Cessie*	Correlação entre resíduos, Somatório dos resíduos quadráticos alisados Ideal: p>0,05	Avalia não-linearidades ou efeitos de interação	Restrita a modelos logísticos e multinomiais. Interpretação difícil.

Modelos de Predição Testados

A tabela 2 descreve o LRM, e a figura 1 descreve o modelo CART. O LRM selecionou cirurgia de emergência (OR = 5,28), DPOC (OR = 4,29), disfunção ventricular (OR = 2,68) e idade (OR = 1,04) como preditores independentes. O modelo CART selecionou adicionalmente angina instável, índice de massa corporal (IMC) baixo e peso como preditores, além dos descritos na LRM, com 11 nós terminais ("folhas") e probabilidades de variação pneumonia entre zero e 66%.

Cirurgia de emergência foi o preditor mais forte em ambos os modelos, aumentando a chance de pneumonia em quase seis vezes, representando, provavelmente, uma *Proxy* para a gravidade da doença e de outros fatores cirúrgicos, como a duração da cirurgia, trauma pulmonar e hemorragia. Na literatura, a DPOC e a idade são normalmente associados com pneumonia. A disfunção ventricular e angina instável são sinais de doença cardíaca grave. Baixo IMC e peso estão associados com doença valvular grave.

Tabela 2 - Modelo de regressão logística

Variável	**Categoria**	**Razão de chances**	**IC 95%**	**P**
DPOC	Sim	4,29	1,73 – 10,61	0,02
	Não	1,00		
Cirurgia de Emergência	Sim	5,58	2,56-12,12	<0,01
	Não	1,00		
Idade	Ordinal	1,04	1,0 - 1,08	0,02
Disfunção ventricular moderada/grave	Sim	2,68	1,27- 5,66	0,01
	Não	1,00		

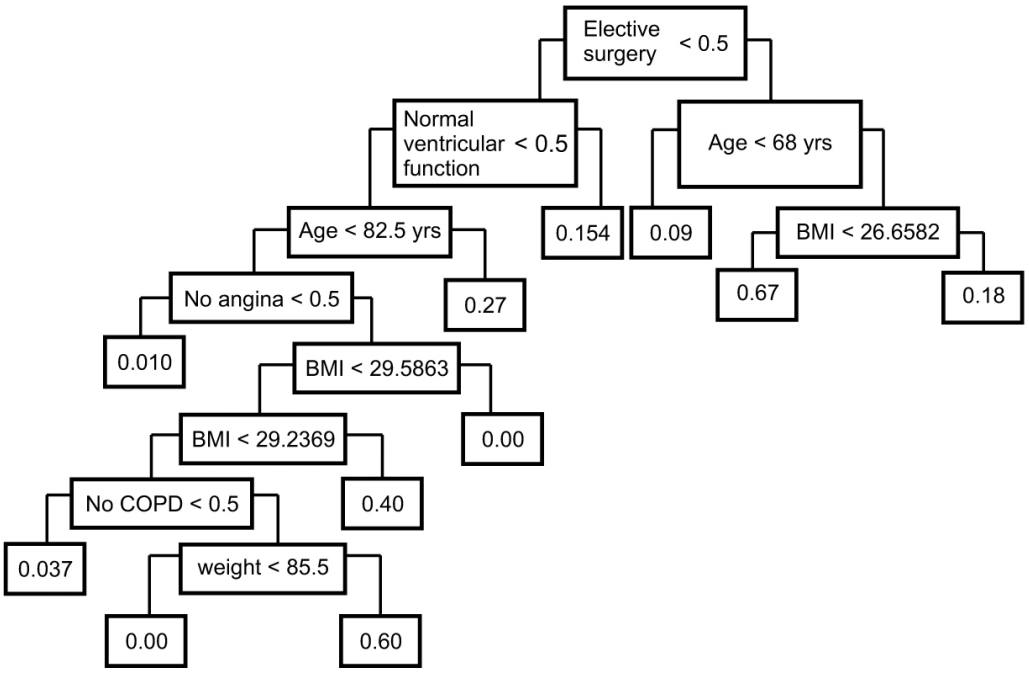

Figura 1 - Modelo de ärvore de classificação

Resultados

A amostra de validação da coorte foi composta por adultos de meia-idade, principalmente do sexo masculino. A coorte de validação externa teve uma baixa percentagem de cirurgia de emergência, de baixo valor médio de creatinina, de alta prevalência de disfunção ventricular importante, e maior tempo de internação antes da cirurgia quando comparados à coorte de derivação (Tabela 3).

A incidência de pneumonia na coorte de validação foi comparável à da coorte de derivação (7,6%). A pneumonia foi diagnosticada em presença de disfunção respiratória progressiva em 74% dos casos de pneumonia e esta característica foi observada em apenas 8,3% dos pacientes sem pneumonia. Nesta amostra, a confirmação microbiológica estava disponível em 61% dos casos. As enterobactérias predominaram entre os patógenos isolados. Febre estava presente em 72% dos casos, a tosse em 38%, estertores em 43%, leucocitose em 87%, estado mental alterado em 32%, e a mudança no caráter da expectoração em 76%. O LRM apresentou desempenho melhor do que o modelo CART nas medidas globais, discriminação e de calibração. O R^2 mostrou que apenas uma pequena quantidade da

variação nos dados é explicada por esses modelos, e o escore de *Brier* foi baixo e semelhante entre os dois modelos (Tabela 4)

Tabela 3 - Características da população da amostra de validação

Variável	**Características da Coorte de Validação**
Número de Casos	333
Idade média em anos (intervalo)	56 (18-85)
Mulheres (%)	44
Tempo pré-operatório médio em dias (intervalo)	22,7(0-148)
Creatinina (media mg/dl)	0,9
DPOC (%)	4,9
IMC (média Kg/m^2)	25,6
Peso (média Kg)	68,4
Cirurgia de Emergência (%)	1,5
Angina Instável (%)	11,0
Disfunção ventricular Moderada/grave (%)	19
Incidência de Pneumonia (%)	7,5

Tabela 4 - Resultados das medidas de desempenho

		Validação Externa (n=333)	
		LRM	CART
Medidas Globais	R^2	7,1%	4%
	Brier	0,06	0,07
Medidas de Discriminação	indice C	0,694	0,61
	95% IC	0,5851894 - 0,8006548	0,56 - 0,67
	Angulo de Discriminação	0,027	0,045
	95% CI	0,004595652- 0,03385048	0,00-0,09
Medidas de Calibração	Angulo de Calibração	0,77	0,04
	Hosmer-Lemeshow	0,08	p<0,001
	Goeman le Cessie	0,26	–

O LRM apresentou valores de discriminação pelo índice C melhor dos que o de CART, mas sem uma diferença significativa. Considerando o ângulo de discriminação, foi observada uma baixa capacidade discriminativa com pouca superioridade do modelo CART.

Em termos de calibração, encontramos as principais diferenças entre o desempenho dos modelos; o ângulo de calibração do LRM foi próximo de um, e muito maior do que o do modelo CART (Tabela 4). O teste de *Hosmer-Lemeshow* indica uma clara vantagem para a calibração do LRM, confirmada com uma boa calibração para o LRM no teste *Goeman Le Cessie*. Os aspectos do gráfico de calibração do LRM estavam mais perto da linha ideal (a previsão foi semelhante à dos resultados observados) quando comparados aos de CART, mas superestimado para as probabilidades acima de 40% das pneumonias (Figura 2 e 3).

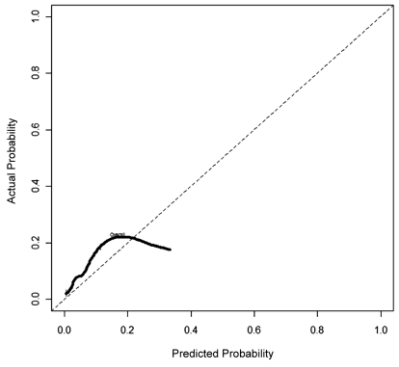

Figura 2 - Gráfico de calibração modelo de regressão logística

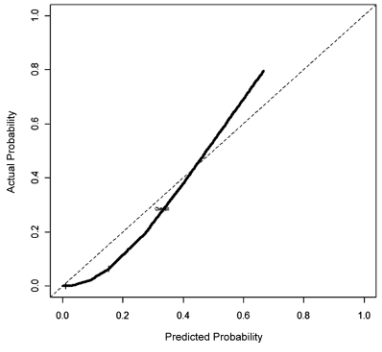

Figura 3 - Gráfico de calibração de árvore de classificação

Discussão

A população do estudo apresentava algumas particularidades sendo de um hospital público terciário no Brasil. A baixa prevalência de cirurgia de emergência é explicada pela ausência de um setor de emergência aberta e um maior uso de intervenção coronária percutânea na Síndrome Coronariana Aguda. Outras características importantes foram o tempo de permanência do pré-operatório, devido, basicamente, aos atrasos nos exames laboratoriais e ao número de cirurgias disponíveis nos hospitais públicos (Tabela 3). A falta de acesso aos centros terciários está associada com uma maior prevalência de pacientes com disfunção ventricular grave. O baixo valor médio da creatinina é devido à organização do sistema de saúde em relação aos pacientes com insuficiência renal pré-operatória que são redirecionados a outro hospital na cidade.

O desempenho comparativo de medidas globais (R^2) apresentou valores baixos, refletindo, provavelmente, as limitações da modelagem de fenômenos complexos como a pneumonia hospitalar. O escore de Brier para ambos os modelos apresentou valores próximos de ideais, considerando-se que valores menores são melhores e que 0,25 correspondem ao acaso.

As medidas de discriminação mostraram resultados comparáveis aos da literatura com índice C com valores variando de 0,52 - 0,72, com apenas um modelo para a previsão de pneumonia após a cirurgia não-cardíaca [10] com um valor índice C de 0,84. A interpretação do índice C depende do problema clínico, por exemplo, o diagnóstico de meningite bacteriana pode ser estabelecido através de um teste único com um índice C acima de 0,80. Para melhorar o teste diagnóstico, um modelo que combina resultados de diversos testes deve alcançar um índice C de pelo menos 0,90. Por outro lado, em situações em que os instrumentos precisos para o tratamento ainda não estão disponíveis, tais como a seleção dos casais para a fertilização intra-uterina, um modelo de predição com um índice C de 0, 65[11] já pode ser útil.

As medidas de calibração exibiram uma grande diferença, como ângulo de calibração da LRM mais próximo de um (valor ideal). Esta medida reflete a curva de calibração onde as freqüências observadas são plotadas em função de probabilidades preditas. Idealmente, o gráfico mostra uma linha de 45 ° com uma inclinação de um. O intercepto e o ângulo da reta de calibração podem ser estimados em uma LRM com o preditor linear do *log* das chances [12]. Um ângulo menor do que um indica que as previsões são muito extremo, muito baixas

para predições pequenas ou muito altas para predições grandes. A não-calibração com o ângulo próximo a um e o intercepto diferente de zero é um achado típico em estudos de validação externa, conforme descrito por Steyerberg at al.[7] "Isso indica que certas características do paciente, que não foram incluídos no modelo de predição, foram diferentemente distribuídas na amostra de validação em comparação com a amostra de derivação. Por exemplo, os pacientes da amostra de validação foram provenientes de um hospital terciário, enquanto que os pacientes da amostra de derivação eram principalmente provenientes de hospitais de referência secundária".

O teste de Hosmer-Lemeshow é obtido pela formação de 10 grupos contendo os decis dos valores ajustados. Valores observados e esperados são calculados [12]. A hipótese nula do teste é um bom ajuste do modelo. Os resultados não mostraram rejeição da hipótese nula para o LRM (P = 0,08) e de rejeição para a CART (P <0,01). O teste de Goeman *le Cessie* é baseado em uma soma dos quadrados residuais, com uma interpretação semelhante à de Hosmer-Lemeshow. Goeman le Cessie foi realizada apenas para LRM por limitações técnicas, sem rejeição da hipótese nula. Um dos resultados de desempenho que merece atenção é quando as diferenças na calibração refletem uma diferença na incidência de desfechos entre o conjunto de derivação e de validação definido que não pode ser explicado por diferentes distribuições dos valores dos preditores. Esses valores podem ser usados para atualizar o modelo nos chamados "métodos de recalibração "[13].

Vários estudos têm mostrado que a validação externa dos modelos preditivos é necessária para garantir sua capacidade de transferência para novos grupos (generalização do modelo)[14]. É essencial a *validação externa* antes de incorporar modelos de predição na prática clínica; modelos de previsão que não são validados normalmente não são preparados para a aplicação clínica. Um modelo submetido a um procedimento de validação externa com bons resultados são considerados nível II, no desenvolvimento do modelo preferencialmente seguido de avaliação de impacto passando a ser nível I.[14].

Quase todos os modelos de predição de infecção não foram submetidos à validação externa e, provavelmente, devem conter otimismo no desempenho como resultado da amostra da derivação. Os modelos são geralmente submetidos apenas à validação interna pelo método de divisão da amostra ("*split*") ou "*boostrap*". Os resultados são geralmente aceitos sem considerar a importância da validação externa [4], está se limitando à generalização de um modelo de predição em novos cenários. A maioria dos modelos de predição no controle de infecção não chega à validação externa[15-23] e podem perder precisão durante a validação. Um pequeno número de estudos avaliou a predição de pneumonia hospitalar como

um alvo, e só um criou uma regra de predição de pneumonia após cirurgia cardíaca, sem validação[19].

Modelos preditivos com objetivos diagnósticos e prognósticos são criados para fins diferentes. Modelos de diagnóstico são utilizados para classificar os indivíduos em seus estados de doença verdadeira, enquanto os modelos prognósticos têm um objetivo mais complexo, estimando-se o risco individual e alocando corretamente o paciente em estratos de riscos diferentes[24]. Avaliação com base apenas em discriminação, usando somente o índice C, não permite quantificar corretamente o desempenho um modelo prognóstico. Uma estatística de calibração avalia a forma como os valores preditos concordam com os valores observados nos dados de validação. Um grande conjunto de medidas de desempenho incorpora aspectos globais, discriminação e calibração podendo auxiliar na comparação entre o impacto clínico dos dois modelos sobre os riscos para o indivíduo, bem como para a população.

Duas principais causas da perda de precisão durante a validação externa são possíveis, incluindo uma menor incidência de fatores preditivos - chave como a cirurgia de emergência. Outro ponto fundamental é o tamanho da amostra. *Harrell et al.*[25] afirma que não há critérios fixos para o cálculo do tamanho da amostra com base em regras de predição, mas em um sentido prático deve-se contar com um mínimo de 10 pacientes com cada fator preditor na amostra de Validação. Morise et al.[26] destaca o impacto das mudanças na prevalência de fatores ao longo do desempenho de um Modelo. *Vergouwe et al.*[27] sugeriu que pelo menos 100 eventos e 100 não-eventos são necessários em uma amostra de validação Externa. *Peek at al.*[13] mostrou que, em modelos de predição de óbito, amostras maiores do que 5.000 pacientes são necessários para obter o Poder de 95% de certeza para a AUC e o escore de *Brier*.

A literatura aponta para a possibilidade de superajuste da árvore de classificação, principalmente por duas razões principais[9]: a regra era insuficientemente desenvolvida ou havia grandes diferenças entre as populações de derivação e de validação. Nenhum dos dois motivos pode ser descartado. A infecção é um evento adverso multicausal e não é bem compreendido até agora. Decisões sobre escolha de variáveis podem levar a erros. Além disso, cirurgia cardíaca está mudando continuamente, especialmente após o desenvolvimento de métodos de revascularização percutânea. A variável mais importante na previsão da amostra de derivação (cirurgia de emergência) foi extremamente reduzida no conjunto de dados de validação. Um dos grandes problemas com modelos de predição são a diferença de tempo entre a derivação, validação e uso clínico, podendo levar a um modelo obsoleto.

Uma limitação da metodologia de árvore de regressão é o fato de que uma árvore superajustada e extremamente complexa não terá um bom desempenho em um novo conjunto de dados. As possíveis vantagens do modelo de predição CART é que ela pode ser mais fácil de interpretar do que modelos de predição com base em regressão logística, com uma apresentação gráfica simples. Há uma série de desvantagens da análise CART, em especial, na aplicação ela é relativamente recente e árvores de classificação nem sempre podem oferecer pontos de corte clinicamente interpretáveis[9].

O presente estudo tem algumas limitações. Uma limitação importante é não incluir variáveis operacionais tais como a duração da cirurgia, dias sob ventilação mecânica, ou reoperação. Contudo, o objetivo primário do estudo foi utilizar um modelo de previsão para planejar medidas de prevenção antes da cirurgia, restringindo o uso de dados pós-operatórios. Na prática clínica, o desenvolvimento de modelos bem calibrados através de diversos cenários, com um número limitado de indicadores pode ser difícil, deixando algumas variações entre os pacientes inexplicáveis[12].

Torna-se difícil o diagnóstico acurado de pneumonia associada à ventilação porque muitas complicações decorrentes de cirurgia cardíaca e de cuidados intensivos, como congestão pulmonar, êmbolos, contusão, e atelectasias, podem causar imagens radiológicas e febre, que imita o aspecto clínico de pneumonia associada à ventilação[28]. O uso de critérios do *CDC* e a limitação do diagnóstico de apenas um médico treinado em doenças infecciosas reduziram a possibilidade de erro diagnóstico.

As variáveis preditivas selecionadas concordam em parte com o modelo de predição apresentado pela *Hortal*[19], ambos os modelos incluindo idade e cirurgia de emergência. Uma grande vantagem do estudo foi à utilização de dados simples que qualquer um pode obter. A utilização das ferramentas epidemiológicas mais avançadas contribui para o desenvolvimento de mais evidências robustas na área de controle de infecção.

Embora a aplicação prática não tenha sido o objetivo do presente trabalho, alguns pontos merecem atenção. *Reilly*[29] lista as principais barreiras à utilização efetiva das regras de predição, especialmente o medo de diminuição da autonomia, a convicção de que o julgamento clínico é superior à regra, à desconfiança na precisão dos preditores da regra ou à tradução das previsões em decisões, fracos incentivos para a utilização de uma regra, a preocupação de que os fatores importantes que não são abordados por uma regra de decisão (por exemplo, a comorbidade do paciente ou a disponibilidade de recursos), o instrumento regra de decisão não é fácil de usar, e a falta de infra-estrutura de apoio. Para a maioria dos entraves, as soluções são estudar o impacto e a utilização de programas de computador para calcular as previsões facilitando o uso na beira de leito.

Referências

[1] Kollef MH, Sharpless L, Vlasnik J, Pasque C, Murphy D, Fraser VJ. The impact of nosocomial infections on patient outcomes following cardiac surgery. Chest. 1997 Sep;112(3):666-75.

[2] Rebollo MH, Bernal JM, Llorca J, Rabasa JM, Revuelta JM. Nosocomial infections in patients having cardiovascular operations: a multivariate analysis of risk factors. The Journal of thoracic and cardiovascular surgery. 1996 Oct;112(4):908-13.

[3] Toll DB, Janssen KJ, Vergouwe Y, Moons KG. Validation, updating and impact of clinical prediction rules: a review. Journal of clinical epidemiology. 2008 Nov;61(11):1085-94.

[4] Bleeker SE, Moll HA, Steyerberg EW, Donders AR, Derksen-Lubsen G, Grobbee DE, et al. External validation is necessary in prediction research: a clinical example. Journal of clinical epidemiology. 2003 Sep;56(9):826-32.

[5] Santos M, Braga JU, Gomes RV, Werneck GL. Predictive factors for pneumonia onset after cardiac surgery in Rio de Janeiro, Brazil. Infect Control Hosp Epidemiol. 2007 Apr;28(4):382-8.

[6] Horan TC, Culver DH, Gaynes RP, Jarvis WR, Edwards JR, Reid CR. Nosocomial infections in surgical patients in the United States, January 1986-June 1992. National Nosocomial Infections Surveillance (NNIS) System. Infect Control Hosp Epidemiol. 1993 Feb;14(2):73-80.

[7] Steyerberg EW. Clinical prediction models : a practical approach to development, validation, and updating. New York ; London: Springer 2009.

[8] Wolfe R, McKenzie DP, Black J, Simpson P, Gabbe BJ, Cameron PA. Models developed by three techniques did not achieve acceptable prediction of binary trauma outcomes. Journal of clinical epidemiology. 2006 Jan;59(1):26-35.

[9] Lewis R. An Introduction to Classification and Regression Tree (CART) Analysis. 2000 Society for Academic Emergency Medicine (SAEM) annual meeting. San Francisco, California, USA. Acad Emerg Med. 2000 May;7(5):419-608.

[10] Arozullah AM, Khuri SF, Henderson WG, Daley J. Development and validation of a multifactorial risk index for predicting postoperative pneumonia after major noncardiac surgery. Annals of internal medicine. 2001 Nov 20;135(10):847-57.

[11] Hosmer DW, Taber S, Lemeshow S. The importance of assessing the fit of logistic regression models: a case study. American journal of public health. 1991 Dec;81(12):1630-5.

[12] Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. Journal of clinical epidemiology. 2008 Jan;61(1):76-86

[13] Peek N, Arts DG, Bosman RJ, van der Voort PH, de Keizer NF. External validation of prognostic models for critically ill patients required substantial sample sizes. Journal of clinical epidemiology. 2007 May;60(5):491-501.

[14] McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-Based Medicine Working Group. Jama. 2000 Jul 5;284(1):79-84.

[15] de Oliveira AC, Ciosak SI, Ferraz EM, Grinbaum RS. Surgical site infection in patients submitted to digestive surgery: risk prediction and the NNIS risk index. American journal of infection control. 2006 May;34(4):201-7.

[16] Escolano S, Golmard JL, Korinek AM, Mallet A. A multi-state model for evolution of intensive care unit patients: prediction of nosocomial infections and deaths. Statistics in medicine. 2000 Dec 30;19(24):3465-82.

[17] Fowler VG, Jr., O'Brien SM, Muhlbaier LH, Corey GR, Ferguson TB, Peterson ED. Clinical predictors of major infections after cardiac surgery. Circulation. 2005 Aug 30;112(9 Suppl):I358-65.

[18] Friedman ND, Russo PL, Bull AL, Richards MJ, Kelly H. Validation of coronary artery bypass graft surgical site infection surveillance data from a statewide surveillance system in Australia. Infect Control Hosp Epidemiol. 2007 Jul;28(7):812-7.

[19] Hortal J, Giannella M, Perez MJ, Barrio JM, Desco M, Bouza E, et al. Incidence and risk factors for ventilator-associated pneumonia after major heart surgery. Intensive care medicine. 2009 Sep;35(9):1518-25.

[20] Lapresta Moros C, Solano Bernad VM, del Villar Belzunce A, Hernandez Navarrete MJ, Gomez-Juarez Sango A, Arribas Llorente JL. [Predictive model for nosocomial pneumonia in intensive care units]. Medicina clinica. 2007 May 26;128(20):761-5.

[21] Shorr AF, Zilberberg MD, Micek ST, Kollef MH. Prediction of infection due to antibiotic-resistant bacteria by select risk factors for health care-associated pneumonia. Archives of internal medicine. 2008 Nov 10;168(20):2205-10.

[22] Staat P, Cucherat M, George M, Lehot JJ, Jegaden O, Andre-Fouet X, et al. Severe morbidity after coronary artery surgery: development and validation of a simple predictive clinical score. European heart journal. 1999 Jul;20(13):960-6.

[23] Zahar JR, Nguile-Makao M, Francais A, Schwebel C, Garrouste-Orgeas M, Goldgran-Toledano D, et al. Predicting the risk of documented ventilator-associated pneumonia for benchmarking: construction and validation of a score. Critical care medicine. 2009 Sep;37(9):2545-51.

[24] Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clinical chemistry. 2008 Jan;54(1):17-23.

[25] Harrell FE, Jr., Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: advantages, problems, and suggested solutions. Cancer treatment reports. 1985 Oct;69(10):1071-77.

[26] Morise AP, Diamond GA, Detrano R, Bobbio M, Gunel E. The effect of disease- prevalence adjustments on the accuracy of a logistic prediction model. Med Decis Making. 1996 Apr-Jun;16(2):133 42.

[27] Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. Journal of clinical epidemiology. 2005 May;58(5):475-83.

[28] Klompas M, Platt R. Ventilator-associated pneumonia--the wrong quality measure for benchmarking. Annals of internal medicine. 2007 Dec 4;147(11):803-5.

[29] Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Annals of internal medicine. 2006 Feb 7;144(3):2019.

6 RESULTADOS COMPLEMENTARES

6.1 Comparação entre os modelos de derivação e na validação segundo análise gráfica

Curva ROC: as figuras 3 e 4 demonstram que a árvore apresentava um modelo superajustado que apresentou grandes perdas na capacidade discriminativa quando submetida à validação externa, em especial, acima do ponto de 0,4 de sensibilidade, onde para pequenos ganhos de sensibilidade ocorrem grandes perdas de especificidade. Em relação ao modelo logístico a perda é menos pronunciada e a curva apresenta-se mais uniforme.

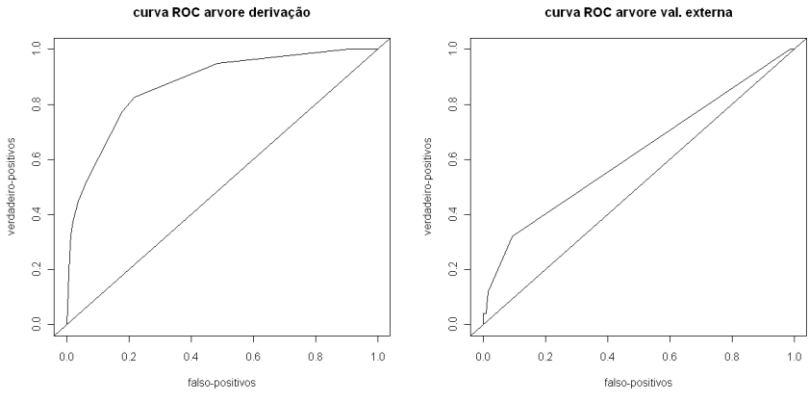

Figura 3 - Curva ROC Modelo Árvore - comparação entre derivação e validação externa

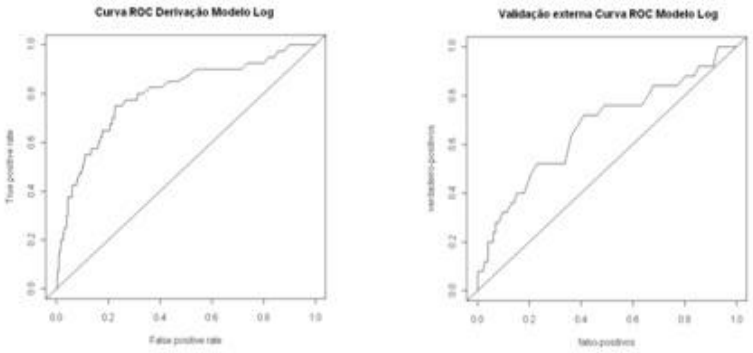

Figura 4 – Curva ROC Modelo de Regressão Logística – comparação entre derivação e validação externa

6.2 **Gráfico de Discriminação**

Em relação ao gráfico de discriminação, o modelo árvore apresenta na derivação um desempenho mais expressivo do que o LRM, sem superposição das caixas, mostrando a capacidade de o modelo separar corretamente os indivíduos com e sem pneumonia. Também se nota no modelo logístico que a baixa incidência do desfecho "achatou" as caixas, e que as predições foram baixas tanto nos pacientes com como os sem pneumonia. Do ponto de vista gráfico, o desempenho do modelo logístico na validação foi prejudicado pela incidência do desfecho, com pequena diferença entre as caixas dos pacientes com e sem pneumonia.

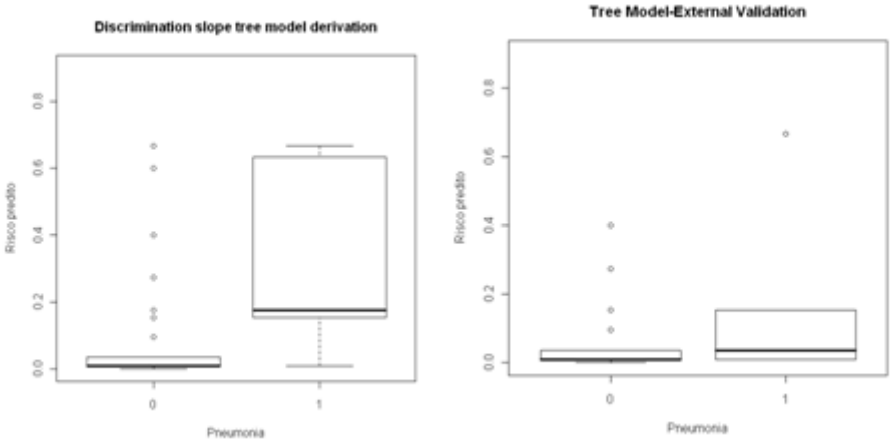

Figura 5 - Gráfico de Discriminação Modelo Árvore - comparação entre derivação e validação externa

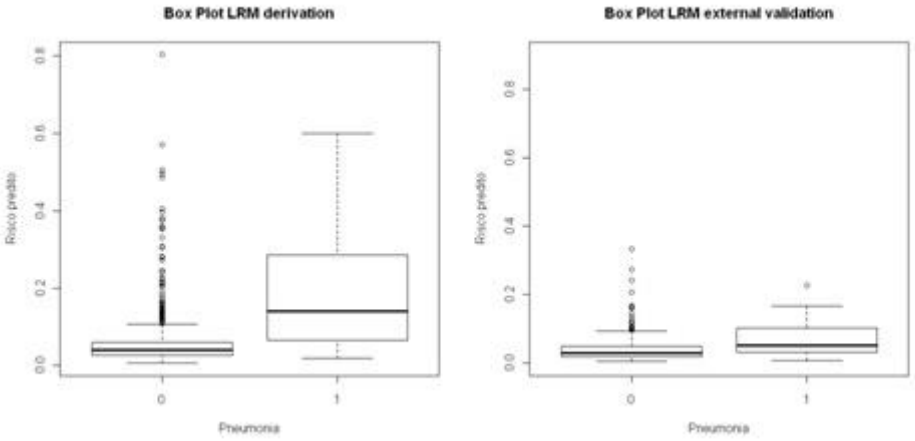

Figura 6 - Gráfico de Discriminação- Modelo de Regressão Logística – comparação entre derivação e validação externa

6.3 **Gráfico de calibração**

A apresentação serve apenas para ilustrar o fato que as "curvas" de calibração na derivação seguem a reta ideal e apenas na validação externa o desempenho real do modelo pode ser avaliado. O modelo árvore concentrou as predições em baixas probabilidades de desfecho. O modelo logístico só foi capaz de predizer baixas probabilidades e que o modelo árvore superestimou as probabilidades acima de 0,2.

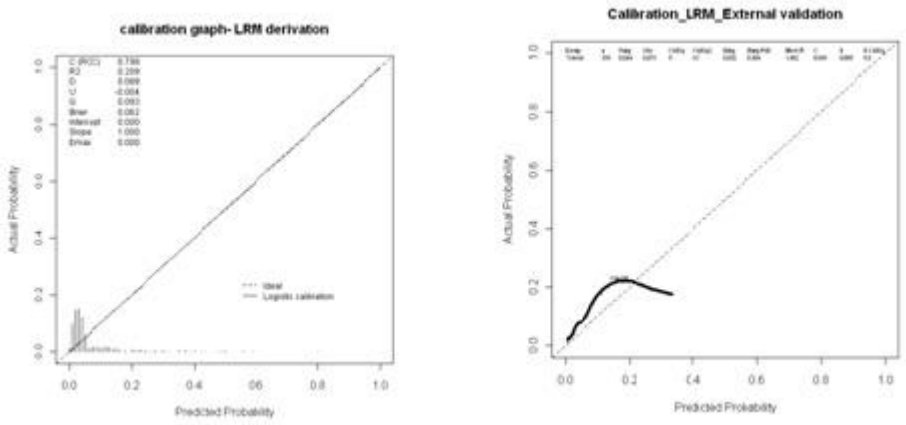

Figura 7 - Gráfico de Calibração Modelo Árvore - comparação entre derivação e validação externa

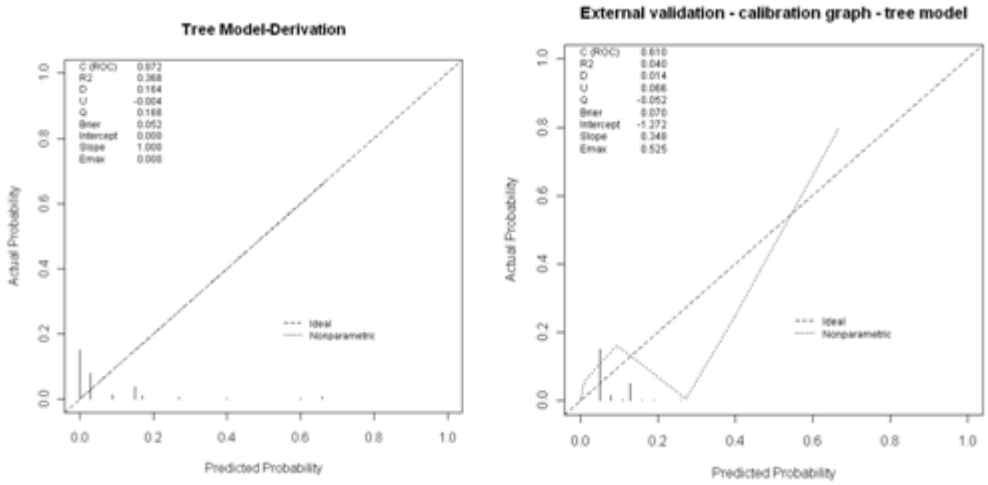

Figura 8 - Gráfico de Calibração Modelo de Regressão Logística – comparação entre derivação e validação externa

7 CONSIDERAÇÕES FINAIS

Grande parte dos modelos na área de infecção ainda não apresenta níveis ideais de desenvolvimento que permitam sua aplicação prática e podem evoluir com perdas na sua acurácia durante a validação. Estas perdas se devem a alguns fatores dentre eles mudanças na população, prática clínica e tratamentos. Outros problemas apontados pela literatura são os erros de classificação e na coleta de dados.

A ausência de variáveis per e pós-operatórias, como o tempo de duração da cirurgia e da ventilação mecânica, pode ter reduzido a acurácia dos modelos. Entretanto, como a intenção deste estudo era a classificação dos indivíduos em grupos de intervenção prioritária para prevenção de pneumonia ainda no estágio pré-operatório, estes dados não poderiam ser incorporados. Mesmo com esta restrição, pode-se obter um sistema preditivo relativamente acurado a partir de combinação de variáveis de simples obtenção, como presença de doença pulmonar obstrutiva crônica (DPOC), idade avançada, cirurgia de emergência e a disfunção ventricular. Aplicação de técnicas epidemiológicas mais elaboradas contribui para o desenvolvimento de evidencias mais robustas na área de controle de infecções hospitalares.

Dois fatores podem ter prejudicado a estratégia de validação: menor incidência de fatores associados ao desfecho, como cirurgias de emergência e o tamanho amostral. Segundo *Harrel* [16] "... não existem critérios fixos para cálculos de tamanhos amostrais em RPC, mas considera-se uma regra prática da presença mínima de dez pacientes com cada variável preditora importante...". *Morise* [17] também destaca o impacto da mudança na prevalência das variáveis no desempenho do modelo. *Vergouwe* [18] sugere um mínimo de 100 eventos e 100 não eventos em uma amostra para validação externa. *Peek* [19] demonstrou que, em modelos de predição de óbito, amostras são ainda maiores com até 5.000 observações para se obter um cálculo com 95% de certeza para obtenção de valores corretos da AUC e do escore de *Brier*.

Em relação à aplicação prática deste modelo alguns pontos merecem destaque, embora não sendo objeto primário deste estudo. *Reilly* [20] listou as principais barreiras para adoção de uma RPC, destacando-se no caso deste estudo:

- Medo da redução da autonomia do médico
- Confiança excessiva no julgamento clínico
- Não confiar na acurácia das predições

- Medo de problemas legais
- Poucos incentivos para adoção de novas regras
- Acreditar que variáveis importantes não foram incluídas
- Medo de prejudicar a segurança do paciente
- Achar que a RPC não é de fácil uso
- Falta de infraestrutura para o uso da RPC
- Tendência natural dos médicos a retornarem a comportamentos anteriores

Para a grande maioria destes itens, as soluções são a realização de estudos de impacto de utilização da regra e a adoção de *softwares* que permitam o cálculo automático da probabilidade predita de pneumonia no momento do agendamento da cirurgia, itens que devem ser planejados de modo a seguir como parte do desenvolvimento desta RPC.

A análise de desempenho dos modelos baseada apenas em dados de derivação ou de validação interna deve ser considerada como otimista e utilizada apenas como base para novos estudos. A incorporação de modelos preditivos deve ser precedida por uma estratégia de validação externa, preferencialmente, em mais de um estudo para garantir a capacidade de generalização dos modelos.

O estudo de técnicas e métodos de modelagem ainda encontra-se em estagio inicial de desenvolvimento, devendo ser estimulada como linha de pesquisa, em especial no campo da cirurgia cardíaca.

O trabalho apresentado criou a oportunidade de revisar um grande número de técnicas de avaliação de desempenho e criar um roteiro para análise utilizando um programa estatístico livre e disponível para uso de outros pesquisadores da área.

REFERÊNCIAS

[1] Brasil, Saude Md, Saude SdAa, Controle DdRAe. Manual técnico do Sistema de Informação Hospitalar / Ministério da Saúde, Secretaria de Atenção à Saúde, Departamento de Regulação, Avaliação e Controle. Série A Normas e Manuais Técnicos. Brasília: Editora do Ministério da Saúde 2007.

[2] Eagle KA, Guyton RA, Davidoff R, Ewy GA, Fonger J, Gardner TJ, et al. ACC/AHA guidelines for coronary artery bypass graft surgery: executive summary and recommendations : A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee to revise the 1991 guidelines for coronary artery bypass graft surgery). Circulation. 1999 Sep 28; 100(13):1464-80.

[3] Giomarelli P, Scolletta S, Borrelli E, Biagioli B. Myocardial and lung injury after cardiopulmonary bypass: role of interleukin (IL)-10. The Annals of thoracic surgery. 2003 Jul; 76(1):117-23.

[4] Drakulovic MB, Torres A, Bauer TT, Nicolas JM, Nogue S, Ferrer M. Supine body position as a risk factor for nosocomial pneumonia in mechanically ventilated patients: a randomised trial. Lancet. 1999 Nov 27; 354(9193):1851-8.

[5] Asensio A, Torres J. Quantifying excess length of postoperative stay attributable to infections: a comparison of methods. J Clin Epidemiol.1999 Dec; 52(12):1249-56.

[6] Rebollo MH, Bernal JM, Llorca J, Rabasa JM, Revuelta JM. Nosocomial infections in patients having cardiovascular operations: a multivariate analysis of risk factors. J Thorac Cardiovasc Surg. 1996 Oct; 112(4):908-13.

[7] Murphy DM. From expert data collectors to interventionists: changing the focus for infection control professionals. Am J Infect Control. 2002 Apr; 30(2):120-32.

[8] Muscedere JG, Martin CM, Heyland DK.The impact of ventilator-associated pneumonia on the Canadian health care system. J Crit Care. 2008 Mar; 23(1):5-10.

[9] Santos M, Braga JU, Gomes RV, Werneck GL. Predictive factors for pneumonia onset after cardiac surgery in Rio de Janeiro, Brazil. Infect Control Hosp Epidemiol. 2007 Apr; 28(4):382-8.

[10] McGinn T, P. W, Wisnisvesky J, Devreaux P, Stiell I, Richardson S. Clinical Prediction Rules. In: Guyatt G, ed. *Users' guides to the medical literature : a manual for evidence-based clinical practice*. 2nd ed. New York: McGraw-Hill Medical 2008:xxiii, 836 p.

[11] Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. Jama. 1997 Feb 12;277(6):488-94.

[12] Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. N Engl J Med. 1985 Sep 26 ; 313(13):793-9.

[13] McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-Based Medicine Working Group. Jama. 2000 Jul 5; 284(1):79-84.

[14] Guyatt G. Users' guides to the medical literature : essentials of evidence-based clinical practice. 2nd ed. New York: McGraw-Hill Medical 2008.

[15] Breiman L. Classification and regression trees. Belmont, Calif.: Wadsworth International Group 1984.

[16 Harrell FE, Jr., Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: advantages, problems, and suggested solutions. Cancer treatment reports. 1985 Oct; 69(10):1071-77.

[17] Morise AP, Diamond GA, Detrano R, Bobbio M, Gunel E. The effect of disease-prevalence adjustments on the accuracy of a logistic prediction model. Med Decis Making.1996 Apr-Jun; 16(2):133-42.

[18] Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol. 2005 May; 58(5):475-83.

[19] Peek N, Arts DG, Bosman RJ, van der Voort PH, de Keizer NF. External validation of prognostic models for critically ill patients required substantial sample sizes. J Clin Epidemiol. 2007 May; 60(5):491-501.

[20] Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Annals of internal medicine. 2006 Feb 7;144(3):2019.

ANEXO A - Rotina em R criada para tese de doutorado

```
require(Design)
require(foreign)
require(tree)
require(ROCR)
require(Hmisc)
require(epiR)
setwd("banco de dados") #fix place for find data
DEVPRO<-read.dta("banco_derivacaoPRO_21nov.dta") #reading stata data
#tree building
#complex tree
arvore_novapro<-
tree(factor(vapn)~age+imc+angins+peso+COPD+ventrdysf+emerg,x=T,y=T,data=DEV
PRO)
plot(arvore_novapro)
text(arvore_novapro)
#opening new window
x11( )
#pruning tree
arv1<-snip.tree(tree = arvore_novapro, nodes = c(5,6,9,16,69,136))
plot(arv1)
text(arv1)
arv1
#finding tree probabilities
DEVPRO$prob_arv<-as.numeric(predict(arv1,type="vector")[,2]+0.00001)
DEVPRO$clas_arv<-predict(arv1,type="class")
DEVPRO$linear.predictor<- as.numeric(qlogis(DEVPRO$prob_arv))

#model diagnosis
val.prob(logit=DEVPRO$linear.predictor,logistic.cal=F, y=DEVPRO$vapn)
x<-val.prob(logit=DEVPRO$linear.predictor, y=DEVPRO$vapn, group=T)
title(paste("Tree Model-Derivation"))
#CI C index
cstat.dev <- rcorr.cens(DEVPRO$prob_arv, DEVPRO$vapn)
cstat.dev
cat("C index:",cstat.dev[1], "[", cstat.dev[1]-1.96/2*cstat.dev[3], " - ",
cstat.dev[1]+1.96/2*cstat.dev[3],"]\n")

#box plot
k3       <-       mean(DEVPRO$prob_arv[DEVPRO$vapn==1])       -       mean
(DEVPRO$prob_arv[DEVPRO$vapn==0])
k3
boxplot(DEVPRO$prob_arv~          DEVPRO$vapn,ylab="Risco          predito",
xlab="Pneumonia",ylim=c(0,0.9))
title(paste("Tree Model-Derivation"))


#Bootstrap for CI discrimination slope
nrowB.arv   <- nrow(DEVPRO)  # number of coluns equal to derivation
B <- 200            # 200 bootstraps
matriz.B.arvore <- matrix(NA,nrow=B,ncol=3) # Matriz for results
dimnames(matriz.B.arvore)    <-    list(c(1:B),    Cs(Slopeapp,    Slopetest,
optimism ))#names for variables
# Start loop
for (i in 1:B) { if (i%%10==0) cat("Start Bootstrap sample nr", i, "\n")
Brows.arv <- sample(nrowB.arv,replace=T) #sample with repositioning
Bsample.arv<- DEVPRO[Brows.arv,] #sample with bootstrap from derivation
```

```
prob.nova.arv              <-       predict.tree(object=arv1,       newdata=
Bsample.arv,type="vector")[,2]#putting Bsample into the tree model
matriz.B.arvore[i,1]    <-    mean(prob.nova.arv[Bsample.arv$vapn==1])    -
mean(prob.nova.arv[Bsample.arv$vapn==0])    #discrimination    slope    de
deiscriminaçao na Bsample
matriz.B.arvore[i,2]    <-    mean(prob.nova.arv[arvore_novapro$y==1])    -
mean(prob.nova.arv[arvore_novapro$y==0])  # testing into the priginal data
set
} #End for loop
cat("\n\n\n\n")
matriz.B.arvore[,3]  <-  matriz.B.arvore[,1]  -  matriz.B.arvore[,2]  #
calculating optmism
# matB mean en SD
apply(matriz.B.arvore,2,mean)
apply(matriz.B.arvore,2,sd)
k1.arv <- apply(matriz.B.arvore,2,mean)
k2.arv <- apply(matriz.B.arvore,2,sd)
cat("CI: ",k3, "[", k3-1.96*k2.arv[2], " - ", k3+1.96*k2.arv[2],"]\n")#
calculating the CI

#Hosmer Lemeshow
source("HLtest.r") #call for macro of HL test
hl.ext2(p=DEVPRO$prob_arv,y=DEVPRO$vapn,g=10,df=8)

#External Validation
valid<-read.dta("banco_validacaoINCi.dta") #reading the external dataset
DEVPRO.lrm$y<-as.numeric(DEVPRO.lrm$y)
valid$prob_arv<-as.numeric(predict(arv1,newdata=valid,type="vector")[,2]) #
extracting the predict probabilities
valid$clas_arv<-as.numeric(predict(arv1,newdata=valid,type="class"))
valid$linear.predictor<- as.numeric(qlogis(valid$prob_arv)) #extracting the
linear predictors
#diagnosis
val.prob(logit=valid$linear.predictor,logistic.cal=F,y=valid$vap)
#calibration graph
title(paste("External validation - calibration graph - tree model"))
#C-6 CI ROC
castat.val <- 0.612
cat(castat.val,    "[",    castat.val-1.96/2*cstat.dev[3],    "    -    ",
castat.val+1.96/2*cstat.dev[3],"]\n")
valid<-valid[!is.na(valid$vap),]


#C-7 BOX PLOT and DISCR SLOPE#
boxplot(plogis(valid$linear.predictor)~    valid$vap,ylab="Predict    Risk",
xlab="Pneumonia",ylim=c(0,0.9)) #building boxplot
discrimination.slope  <-  mean(plogis(valid$linear.predictor[valid$vap==1]))
-      mean(plogis(valid$linear.predictor[valid$vap==0]))#      calculating
discrimination slope
discrimination.slope
title(paste("Tree Model-External Validation"))

#C-8 CI DISCR slope uses the bootstrap error
cat(discrimination.slope, "[", discrimination.slope-1.96*k2.arv[2], " - ",
discrimination.slope+1.96*k2.arv[2],"]\n")
#hl
hl.ext2(p=valid$prob_arv,y=valid$vap,g=10,df=8)
#LRM
# Derivation Model
#building model
```

```
DEVPRO.lrm <- lrm(vap ~ age+COPD+emerg+ventrdysf, data=DEVPRO,
x=T,y=T,linear.predictors=T)
#extracting linear predictors
lp.lrm <- DEVPRO.lrm$linear.predictors
#diagnosis
DEVPRO.lrm$y <- as.numeric(DEVPRO.lrm$y)
val.prob(logit=DEVPRO.lrm$linear.predictors,smooth=F, y=DEVPRO.lrm$y)
#model diagnosis
title(paste("calibration graph- LRM derivation"))
#calibration diagnosis
LR<-val.prob(logit=DEVPRO.lrm$linear.predictor, group=T, y=DEVPRO.lrm$y)
plot(LR)
#6-CI C index
cstat.dev.lrm <- rcorr.cens(DEVPRO.lrm$linear.predictors, DEVPRO.lrm$y)
cat(cstat.dev.lrm[1], "[", cstat.dev.lrm[1]-1.96/2*cstat.dev.lrm[3], " - ",
cstat.dev.lrm[1]+1.96/2*cstat.dev.lrm[3],"]")

#7-discrimination slope and boxplot
mean(plogis(lp.lrm[DEVPRO.lrm$y==1])) - mean
(plogis(lp.lrm[DEVPRO.lrm$y==0]))# calculanting slope
boxplot(plogis(lp.lrm)~ DEVPRO.lrm$y,ylab="Risco predito", xlab="Pneumonia")
title(paste("Box Plot LRM derivation"))

#8-calculating sd #erro
# Bootstrap discrimination slope
nrowB <- nrow(DEVPRO)  # nrow from development set
B <- 200            # 200 bootstraps
matB <- matrix(NA,nrow=B,ncol=3) # Matrix for results
dimnames(matB) <- list(c(1:B), Cs(Slopeapp, Slopetest, optimism ))
# Start loop
for (i in 1:B) {
if (i%%10==0) cat("Start Bootstrap sample nr", i, "\n")
Brows <- sample(nrowB,replace=T)

# Bsample is bootstrap sample from development set
Bsample     <- DEVPRO[Brows,]
devfull.lrm       <-       lrm(vap       ~       age+COPD+emerg+ventrdysf,
data=Bsample,linear.predictors=T, x=T, y=T)
matB[i,1]  <-  mean(plogis(devfull.lrm$linear.predictors[devfull.lrm$y==1]))
-
            mean(plogis(devfull.lrm$linear.predictors[devfull.lrm$y==0]))
lp.b.lrm   <- DEVPRO.lrm$x %*% devfull.lrm$coef[2:length(devfull.lrm$coef)]
+ devfull.lrm$coef[1] # lp with coefs from bootstrap
matB[i,2]        <-        mean(plogis(lp.b.lrm[DEVPRO.lrm$y==1]))        -
mean(plogis(lp.b.lrm[DEVPRO.lrm$y==0]))  # Testing on original sample
matB[i,2]
} # End for loop
cat("\n\n\n\n")
matB[,3] <- matB[,1] - matB[,2] # optimism
apply(matB,2,mean)
apply(matB,2,sd)
k1 <- apply(matB,2,mean)
k2 <- apply(matB,2,sd)

#5-CI discrimination slope
k3          <-mean(plogis(lp.lrm[DEVPRO.lrm$y==1]))          -          mean
(plogis(lp.lrm[DEVPRO.lrm$y==0]))# calculando a slope
cat(k3, "[", k3-1.96*k2[2], " - ", k3+1.96*k2[2],"]")

#8-H-L
DEVPRO.lrm$y<- as.numeric(DEVPRO.lrm$y)
```

```
hl.ext2(p= plogis(DEVPRO.lrm$linear.predictor),y=DEVPRO.lrm$y,g=10,df=8)

#9-Goeman le Cessie
source("multinomialcode.R") # call for code
lform <- vap ~ age+COPD+emerg+ventrdysf #including the model
fit <- mlogit(lform, DEVPRO) #multinomial
U <- smoothU(DEVPRO, c("age", "COPD","emerg","ventrdysf"),
0.25,"k")#smoothing of residuals
testfit(fit, U %*% t(U)) #output of test, first value = p value (ideal>0.05)

#Internal validation
DEVPRO.lrm <- lrm(vap ~ age+COPD+emerg+ventrdysf, data=DEVPRO,
x=T,y=T,linear.predictors=T)
val.int <- validate(DEVPRO.lrm, B=200) #internal validation
val.int
castat.val <- (val.int[1,5]*0.5)+0.5
cat(castat.val, "[", castat.val-1.96/2*cstat.dev[3], " - ",
castat.val+1.96/2*cstat.dev[3],"]")
 #creating  nomogram
setwd("C:/Users/Marisa/Documents/doutorado1/bancos")
nomo<-read.dta("banco_derivacaoPRO.dta")
attach(nomo)
ddist<-datadist(age,ventrdysf,emerg,COPD)
options(datadist='ddist')
f<-lrm(vapn~ age+ventrdysf+emerg+COPD)
nomogram(f,fun=plogis,funlabel="Pneumonia Risk", lp=F, xfrac=.25)
# External Validation
#2-extracting lienar predictors
lp.ext.lrm  <- predict(object=DEVPRO.lrm, newdata= valid)
#3-diagnosis
val.prob(logit=lp.ext.lrm, y=valid$vap)
LRE<-val.prob(logit=lp.ext.lrm,group=T,smooth=F,y=valid$vap)
plot(LRE)
title(paste("Calibration_LRM_External validation"))
#6- CI C index
cstat.ext.lrm <- rcorr.cens(lp.ext.lrm , valid$vap)
cstat.ext.lrm
cat(cstat.ext.lrm[1], "[", cstat.ext.lrm[1]-1.96/2*cstat.ext.lrm[3], " - ",
cstat.ext.lrm[1]+1.96/2*cstat.ext.lrm[3],"]\n")
#7-boxplot and dicrimination slope
mean(plogis(lp.ext.lrm[valid$vap==1]))-
mean(plogis(lp.ext.lrm[valid$vap==0]))
boxplot(plogis(lp.ext.lrm) ~ valid$vap,ylab="Predict Risk",
xlab="Pneumonia",ylim=c(0,0.9))
title(paste("Box Plot LRM external validation"))
# Bootstrap discrimination slope
nrowB.v <- nrow(valid)  # nrow from development set
B <- 200            # 200 bootstraps
matB.v <- matrix(NA,nrow=B,ncol=3) # Matrix fovalidr results
dimnames(matB.v) <- list(c(1:B), Cs(Slopeapp, Slopetest, optimism ))
# Start loop
for (i in 1:B) {
if (i%%10==0) cat("Start Bootstrap sample nr", i, "\n")
Brows <- sample(nrowB.v,replace=T)
# Bsample is bootstrap sample from development set
Bsample.v<- valid[Brows,]
devfull.v <- lrm(vap ~ age+COPD+emerg+ventrdysf , data=Bsample.v,
linear.predictors=T, x=T, y=T)
matB.v[i,1] <- mean(plogis(devfull.v$linear.predictors[devfull.v$y==1])) -
mean(plogis(devfull.v$linear.predictors[devfull.v$y==0]))
```

```
lp.v    <-  devfull.v  $x  %*%  devfull.v$coef[2:length(devfull.v$coef)]  +
devfull.v$coef[1] # lp with coefs from bootstrap
matB.v[i,2]       <-       mean(plogis(lp.ext.lrm[devfull.v$y==1]))       -
mean(plogis(lp.ext.lrm[devfull.v$y==0]))# Testing on original sample
} # End for loop
cat("\n\n\n")
matB.v[,3] <- matB.v[,1] - matB.v[,2] # optimism
 apply(matB.v,2,mean)
apply(matB.v,2,sd)
k1.v <- apply(matB.v,2,mean)
k2.v <- apply(matB.v,2,sd)
k4.v <- mean(plogis(lp.v[valid$vap==1])) - mean (plogis(lp.v[valid$vap==0]))
cat(k4.v, "[", k4.v-1.96*k2.v[2], " - ", k4.v+1.96*k2.v[2],"]")
```

ANEXO B - Rotina para Teste Hosmer-Lemeshow criado por Yvonne Vergouwe e Ewout Steyerberg

```
# function to perform Hosmer-Lemeshow test for external validation
# instead of chi square and corresponding p-value, this function provides
the number of subjects per group,
# and the mean values of p and y per group.
##########################
# p   : predicted probability
# Y   : outcome variable
# g   : number of groups to calculate H-L (10 is default)
#
# NB: the library Hmisc need to be attached in order to be able to
run hl.ext2
##########################
hl.ext2<-function(p,y,g=10, df=g-1)
{
matres      <-matrix(NA,nrow=g,ncol=5)
sor   <-order(p)
p     <-p[sor]
y     <-y[sor]
groep <-cut2(p,g=g)                                                    #g
more or less equal sized groups
len         <-tapply(y,groep,length)                                   #n      per
group
sump  <-tapply(p,groep,sum)                                  #expected     per
group
sumy  <-tapply(y,groep,sum)                                  #observed     per
group
meanp <-tapply(p,groep,mean)
                                                             #mea
n probability per group
meany <-tapply(y,groep,mean)                                 #mean
                                                             observed per
group
matres      <-cbind(len,meanp,meany, sump, sumy)
contr<-((sumy-sump)^2)/(len*meanp*(1-meanp))         #contribution
                                                     per group to chi
square
chisqr<-sum(contr)
                                                                       #ch
i square total
pval<-1-pchisq(chisqr,df)                                              #p-
value corresponding to chi square with df degrees of freedom
cat("\nChi-square",chisqr," p-value", pval,"\n")
dimnames(matres)  <-list(c(1:g),Cs(n,avg(p),avg(y), Nexp, Nobs))
result  <- list(table(groep), matres,chisqr,pval)
}
```

ANEXO C - Rotina para Teste de Goeman Le Cessie criado por Jelle Goeman

```
#=========================================================
# The class "mlogit.fit" stores the fit of a logistic
# regression model
#=========================================================
setClass("mlogit.fit", representation(
    deviance = "numeric",
    coefficients = "matrix",
    probabilities = "matrix",
    x = "matrix",
    y = "factor"
  )
)

setMethod("show", "mlogit.fit", function(object) {
  cat(paste("deviance =", object@deviance, "\n"))
  cat("coefficients:\n")
  print(object@coefficients)
})

setMethod("residuals", "mlogit.fit", function(object) {
  Y <- sapply(levels(object@y), function(lvl) object@y == lvl )
  Y - fit@probabilities
})


setReference <- function(fit, out) {
  if (!is(fit, "mlogit.fit")) stop("fit should be a multinomial logistic
regression fit object")
  if (out %in% colnames(fit@coefficients)) {
    fit@coefficients <- sweep(fit@coefficients, 1, fit@coefficients[,out])
  } else {
    stop("out not among outcome categories")
  }
  fit
}

mlogit <- function(formula, data, reference, maxiter = 25, epsilon = 1e-8)
{
  # extract X from formula and data
  if (!is(formula, "formula")) stop("formula should be a formula object")
  if (!is.data.frame(data)) stop("data should be a data.frame")
  n <- nrow(data)
  dummyform <- as.formula(paste("rep(0,n) ~", formula[3]))
  dummyfit <- glm(dummyform, data = data, x = TRUE)
  X <- dummyfit$x
  p <- ncol(X)

  # extract Y from formula and data
  Y <- factor(data[[all.vars(formula)[1]]])
  outs <- levels(Y)
  g <- length(outs)

  # enlarge
  bigY <- as.numeric(as.vector(sapply(outs, function(out) { Y == out })))
  bigX <- diag(g) %x% as.matrix(X)

  # fit parameters
```

```
  # 1: starting values:
  biggamma <- rep(0, p*g)
  iter <- 0
  olddev <- -Inf
  finished <- FALSE

  # 2: Newton-Rhaphson algorithm
  while (!finished) {
    numerator <- exp(bigX %*% biggamma)
    denominator <- rep(rowSums(matrix(numerator, n, g)), g)
    bigmu <- as.vector(numerator / denominator)
    bigW <- diag(bigmu) - outer(bigmu, bigmu) * (matrix(1,g,g) %x% diag(n))
    XWX <- t(bigX) %*% bigW %*% bigX
    #XWX2 <- t( bigX - sweep(matrix(1,g,g) %x% as.matrix(X), 2, bigmu,
"*") ) %*% sweep(bigX, 2, bigmu, "*") ???
    eigs <- eigen(XWX, symmetric = TRUE)
    MP.eigs <- c(1/eigs$values[1:(p*(g-1))], rep(0,p))
    XWX.MP <- eigs$vectors %*% diag(MP.eigs) %*% t(eigs$vectors)
    biggamma <- biggamma + XWX.MP %*% (t(bigX) %*% (bigY - bigmu))
    iter <- iter + 1
    dev <- sum(log(bigmu) * bigY)
    finished <- ( abs(dev - olddev) / (abs(dev) + 0.1) < epsilon ) |
(iter >= maxiter)
    olddev <- dev
  }
  if (iter == maxiter)

  # Calculate mu and W based on final estimates
  numerator <- exp(bigX %*% biggamma)
  denominator <- rep(rowSums(matrix(numerator, n, g)), g)
  bigmu <- as.vector(numerator / denominator)

  probs <- matrix(bigmu, n, g)
  rownames(probs) <- names(Y)
  colnames(probs) <- outs
  coefs <- matrix(biggamma, p, g)
  rownames(coefs) <- colnames(X)
  colnames(coefs) <- outs

  fit <- new("mlogit.fit",
    deviance = dev,
    coefficients = coefs,
    probabilities = probs,
    x = X,
    y = Y
  )

  if (!missing(reference)) fit <- setReference(fit, reference)

  fit
}

testfit <- function(fit, R) {
  if (!is(fit, "mlogit.fit")) stop("fit should be a multinomial logistic
regression fit object")

  outs <- colnames(fit@coefficients)
  g <- length(outs)
  n <- nrow(fit@x)
  findout <- rep(outs, rep(n,g))
```

```
  # check input of R
  if (missing(R))
    R <- make.R(as.data.frame(fit@x[,-1]))
  if ( nrow(R) == n & ncol(R) == n )
    bigR <- diag(g) %x% R
  else
    if ( nrow(R) == n*g & ncol(R) == n*g )
      bigR <- R
    else
      stop("wrong dimension of R")


  p <- ncol(fit@x)
  bigX <- diag(g) %x% as.matrix(fit@x)

  #Calculate mu and W based on final estimates
  bigY  <-  as.numeric(as.vector(sapply(outs,  function(out)  {  fit@y  ==
out })))
  bigmu <- as.vector(fit@probabilities)
  bigW <- diag(bigmu) - outer(bigmu, bigmu) * (matrix(1,g,g) %x% diag(n))

  # Calculate hat matrix to correct for estimation of gamma
  XWX <- t(bigX) %*% bigW %*% bigX
  eigs <- eigen(XWX, symmetric = TRUE)
  MP.eigs <- c(1/eigs$values[1:(p*(g-1))], rep(0,p))
  XWX.MP <- eigs$vectors %*% diag(MP.eigs) %*% t(eigs$vectors)
  IminH <- diag(n*g) - bigW %*% bigX %*% XWX.MP %*% t(bigX)
  tR <- t(IminH) %*% bigR %*% IminH

  # The test statistic
  Q <- (bigY - bigmu) %*% bigR %*% (bigY - bigmu)
  RW <- tR %*% bigW

  EQ <- sum(diag(RW))
  mus <- matrix(bigmu, n, g)

  kappai <- function(stuv) {
    case <- rowSums(outer(outs, stuv, "=="))
    switch(sort(case, decreasing = TRUE)[1],
      { # sort(case) = c(1,1,1,1)
        muis <- mus[,case == 1][,1]
        muit <- mus[,case == 1][,2]
        muiu <- mus[,case == 1][,3]
        muiv <- mus[,case == 1][,4]
        -6*muis*muit*muiu*muiv
      },
      {
      switch(sort(case, decreasing = TRUE)[2],
        { # sort(case) = c(2,1,1,0)
          muis <- mus[,case == 2]
          muit <- mus[,case == 1][,1]
          muiu <- mus[,case == 1][,2]
          2*muis*muit*muiu - 6*muis*muis*muit*muiu
        },
        { # sort(case) = c(2,2,0,0)
          muis <- mus[,case == 2][,1]
          muit <- mus[,case == 2][,2]
          -muis*muit    +    2*muis*muit*muit    +    2*muis*muis*muit    -
6*muis*muis*muit*muit
        })
      },
```

```
          { # sort(case) =
            c(3,1,0,0) muis <-
        mus[,case == 3] muit <-
            mus[,case == 1]
         -muis*muit + 6*muit*muis*muis - 6*muit*muis*muis*muis
       },
           { # sort(case) =
              c(4,0,0,0)
         muis <- mus[,case ==
                   4]
        muis - 7*muis*muis + 12*muis*muis*muis - 6*muis*muis*muis*muis
       }
     )
   }

   varQ <- 2 * sum(diag(RW%*%RW)) + sum(sapply(outs, function(out1) {
     sum(sapply(outs, function(out2) {
       dR12 <- diag(tR[findout == out1, findout == out2])
       sum(sapply(outs, function(out3) {
         sum(sapply(outs, function(out4) {
           sum(dR12  *  diag(tR[findout  ==  out3,  findout  ==  out4])
                     *
kappai(c(out1, out2, out3, out4)))
         }))
       }))
     }))
   }))
   seQ <- sqrt(varQ)
   scl <- varQ / (2 * EQ)
   dfr <- EQ / scl

   p.value <- pf ( (scl * dfr / Q), 10^10, dfr )
   #cat(paste("Q = ", Q, "\n"))
   #cat(paste("EQ =", EQ,"\n"))
   #cat(paste("seQ =", seQ, "\n"))
   #cat(paste("scale =", scl, "\n"))
   #cat(paste("df =", dfr, "\n"))
   #cat(paste("p =", p.value, "\n"))
   c(p.value, Q, EQ, seQ)
}

smoothU  <-  function(data,  variables,  quant,  method  =
          c("neighbours", "kernel")) {
   if (missing(variables)) variables <- names(data)
   if (!is.data.frame(data)) stop("data should be a data.frame")
   if (!all(variables %in% names(data))) stop("incorrect variable
   names") DD <- as.matrix(dist(scale(data[,variables])))
   method <- match.arg(method)
   if ( method == "kernel" ) {
     dDD <- DD[diag(nrow(DD)) == 0]
     h <- quantile(dDD, quant)
     U <- apply(DD, c(1,2), function(x) { as.numeric(abs(x) <= h) } )
   }
   if ( method == "neighbours" ) {
     U <- apply(DD, 1, function(x) { as.numeric(x <= quantile(x, quant)) } )
   }
   sweep(U, 2, colSums(U), "/")
}
```