Comparação de métodos de imputação única e múltipla usando como exemplo um modelo de risco para mortalidade cirúrgica

Comparison of simple and multiple imputation methods using a risk model for surgical mortality as example

Luciana Neves Nunes[I,II]

Mariza Machado Klück[I,III]

Jandyra Maria Guimarães Fachel[I,II]

[I] Programa de Pós-Graduação em Epidemiologia da Faculdade de Medicina da Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil.

[II] Departamento de Estatística do Instituto de Matemática da Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil.

[III] Departamento de Medicina Social da Faculdade de Medicina da Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil.

Correspondência: Luciana Neves Nunes. Departamento de Estatística do Instituto de Matemática, Universidade Federal do Rio Grande do Sul. Av. Bento Gonçalves, 9500, Bairro Agronomia, Porto Alegre, RS CEP: 91509-900. E-mail: lununes@mat.ufrgs.br

Resumo

Introdução: A perda de informações é um problema frequente em estudos realizados na área da Saúde. Na literatura essa perda é chamada de *missing data* ou dados faltantes. Através da imputação dos dados faltantes são criados conjuntos de dados artificialmente completos que podem ser analisados por técnicas estatísticas tradicionais. O objetivo desse artigo foi comparar, em um exemplo baseado em dados reais, a utilização de três técnicas de imputações diferentes. **Método:** Os dados utilizados referem-se a um estudo de desenvolvimento de modelo de risco cirúrgico, sendo que o tamanho da amostra foi de 450 pacientes. Os métodos de imputação empregados foram duas imputações únicas e uma imputação múltipla (IM), e a suposição sobre o mecanismo de não-resposta foi MAR (Missing at Random). **Resultados:** A variável com dados faltantes foi a albumina sérica, com 27,1% de perda. Os modelos obtidos pelas imputações únicas foram semelhantes entre si, mas diferentes dos obtidos com os dados imputados pela IM quanto à inclusão de variáveis nos modelos. **Conclusões:** Os resultados indicam que faz diferença levar em conta a relação da albumina com outras variáveis observadas, pois foram obtidos modelos diferentes nas imputações única e múltipla. A imputação única subestima a variabilidade, gerando intervalos de confiança mais estreitos. É importante se considerar o uso de métodos de imputação quando há dados faltantes, especialmente a IM que leva em conta a variabilidade entre imputações para as estimativas do modelo.

Palavras-chave: Métodos de imputação. Imputação múltipla. Dados faltantes. Não-resposta ao acaso.

Abstract

Introduction: It is common for studies in health to face problems with missing data. Through imputation, complete data sets are built artificially and can be analyzed by traditional statistical analysis. The objective of this paper is to compare three types of imputation based on real data. **Methods:** The data used came from a study on the development of risk models for surgical mortality. The sample size was 450 patients. The imputation methods applied were: two single imputations and one multiple imputation and the assumption was MAR (Missing at Random). **Results:** The variable with missing data was serum albumin with 27.1% of missing rate. The logistic models adjusted by simple imputation were similar, but differed from models obtained by multiple imputation in relation to the inclusion of variables. **Conclusions:** The results indicate that it is important to take into account the relationship of albumin to other variables observed, because different models were obtained in single and multiple imputations. Single imputation underestimates the variability generating narrower confidence intervals. It is important to consider the use of imputation methods when there is missing data, especially multiple imputation that takes into account the variability between imputations for estimates of the model.

Keywords: Imputation methods. Multiple imputation. Missing data. Missing at random.

Introdução

A perda de informações é um problema frequente em estudos realizados na área da Saúde. Sujeitos que não preenchem corretamente um item, pacientes que são perdidos ao longo do estudo ou não-preenchimento de resultados de algum exame são exemplos de possíveis perdas de informação. Na literatura, essa perda é chamada de *missing data* ou dados faltantes[1].

Em situações com dados faltantes, uma abordagem bastante comum é restringir a análise aos sujeitos com dados completos nas variáveis envolvidas. Entretanto, as estimativas obtidas com tais análises podem ser viesadas, especialmente se os indivíduos que são incluídos na análise forem sistematicamente diferentes daqueles que foram excluídos em termos de uma ou mais variáveis[2-4]. Como os métodos de análises estatísticas e aplicativos computacionais foram e são desenvolvidos, em sua maioria, para dados completos, mesmo uma pequena quantidade de dados faltantes pode causar problemas nas estimativas (viés, ineficiência), justificando, então, a necessidade de ser considerado nas análises o problema de dados faltantes[5].

A imputação de dados faltantes tem sido uma estratégia comum para a análise de dados com esse problema. Entende-se por imputação a técnica de preencher os dados faltantes com valores plausíveis. Um atrativo para a utilização de técnicas de imputação é o fato de, após a imputação dos dados, o investigador poder utilizar técnicas tradicionais de análise estatística para dados completos[6-8].

Métodos simples como imputação pela média ou pela mediana, também conhecidos como métodos de imputação única, têm sido bastante usados devido à sua facilidade de implementação. Entretanto, existem desvantagens na utilização desses métodos, tais como a subestimação da variabilidade da variável imputada que gerará intervalos de confiança mais estreitos do que o esperado e a impossibilidade de levar em consideração a variabilidade

que possa existir entre diferentes imputações[1,8,9].

Como alternativa à imputação única e com o objetivo de corrigir suas desvantagens, surgiu na década de 80 a imputação múltipla (IM) proposta por Donald Rubin[2,10]. A ideia da imputação múltipla é a de que cada dado ausente é imputado **m** vezes, gerando **m** bancos de dados completos. Os **m** bancos são analisados separadamente por uma técnica tradicional de análise estatística e finalmente os **m** resultados obtidos são combinados de maneira simples para a análise final[1,6,9].

Embora a imputação múltipla tenha boas propriedades estatísticas, ela ainda não é usada com frequência na área da saúde[1,7,8]. Neste trabalho, que teve como motivação um estudo real sobre o desenvolvimento de um modelo de risco para pacientes submetidos à laparotomia, serão discutidos e comparados dois métodos de imputação única e um método de imputação múltipla.

O uso de modelos de risco que têm por objetivo predizer o curso futuro e desfechos dos processos de doenças tem aumentado muito na área da saúde e é muito importante que eles sejam precisos e confiáveis[11]. Entretanto, este artigo não tem por objetivo discutir os aspectos epidemiológicos do modelo de risco de mortalidade cirúrgica, mas sim divulgar a técnica de imputação múltipla, pois embora a literatura sobre IM tenha tido uma expansão considerável desde o início da década de 90, esse número é grande no que diz respeito a publicações somente com aplicações. Mais textos metodológicos têm que ser desenvolvidos e divulgados para que os pesquisadores possam usar a IM rotineiramente e com confiança[12].

Método

Fonte de dados

O banco de dados utilizado como exemplo neste artigo é composto por variáveis coletadas em prontuários de pacientes internados no Hospital de Clínicas de Porto Alegre (HCPA) no período de fevereiro de 2000 a dezembro de 2002, e que foram submetidos à laparotomia exploratória. Originalmente esse banco de dados foi criado por Klück[13] com o objetivo de desenvolver e validar um escore de risco multifatorial para mortalidade cirúrgica.

O banco de dados era constituído de 651 pacientes, posteriormente separado em duas coortes: a de desenvolvimento e a de validação. Com a coorte de desenvolvimento foi feita a modelagem e com a coorte de validação, o modelo foi validado. No presente artigo foram utilizados na análise os 450 pacientes que no estudo original constituíam a coorte de desenvolvimento, sendo de interesse as seguintes variáveis: óbito (ocorrido em até 30 dias após a realização da cirurgia) e ASA (Avaliação pré-anestésica segundo a *American Society of Anestesiology*) com cinco categorias: I a V (onde V é a mais grave), idade (<75 e ≥75 anos), caráter da cirurgia (eletiva ou de urgência), albumina sérica contínua (para imputação) e categorizada (≤ 2,2; 2,3 a 3,0 e ≥ 3,1 g/dl). Essas variáveis fazem parte do modelo final obtido por Klück[13].

A partir da escolha das variáveis de interesse foram feitos os três tipos de imputações para a albumina, a variável que teve dados faltantes (27,1% de perda). As duas imputações únicas foram feitas da seguinte maneira:

- o valor da variável foi imputado considerando-se o valor da mediana da albumina dos pacientes com dados completos, de acordo com o caráter da cirurgia, isto é, submetidos à cirurgia eletiva (Md=3,1g/dl) ou à cirurgia de urgência (Md=2,4g/dl); e
- imputando-se o valor do limite inferior da faixa de normalidade da albumina sérica (3,5g/dl). O primeiro enfoque foi chamado de "método das medianas" e o segundo de "método do valor normal". Esses métodos são chamados de imputação única porque os valores faltantes são preenchidos uma única vez.

Imputação múltipla

Na década de 1980, Rubin[10] escreveu um livro voltado para a técnica de imputação múltipla (IM) para resolver o problema de não-resposta em pesquisas. Embora a técnica que teoricamente seria melhor que a imputação única tenha surgido há bastante tempo, a IM não pôde ser computacionalmente bem implementada na época, pois para implementá-la foram necessários avanços computacionais, o que só ocorreu mais recentemente. A principal vantagem da IM em relação à imputação única é a de que ela leva em conta a variabilidade entre as imputações nos resultados, enquanto os métodos de imputação única não o fazem, visto ser feita apenas uma imputação para cada dado faltante[6,10]. Ilustrativamente, a IM pode ser representada como na Figura 1.

Na IM, o mais importante é a decisão na primeira etapa, ou seja, a escolha do método de IM que será utilizado para gerar as **m** imputações diferentes, pois é preciso que se avalie o tipo de variável que tem não-resposta e que se leve em conta a relação das observações faltantes com as observações presentes. Também é necessário considerar o mecanismo de ausência e o padrão dos dados faltantes. Os mecanismos de ausência se dividem em: perdas completamente ao acaso (*Missing Completely at Random* - MCAR); perdas ao acaso (*Missing at Random* - MAR) e perdas não-aleatórias (*Not Missing at Random* - NMAR). Os padrões de dados faltantes podem ser monotônicos e não-monotônicos[3,6,10,14,15].

Após as **m** imputações terem sido obtidas no primeiro passo, cada um dos **m** bancos de dados completados pela IM são analisados separadamente por métodos estatísticos tradicionais. Finalmente, as **m** estimativas obtidas podem ser combinadas de maneira simples, como foi proposto por Rubin[10].

O procedimento de combinar as estimativas também é conhecido como as "Regras de Rubin" (*Rubin Rules*) e pode ser usado independentemente do método usado para fazer a IM[6,10].

As regras de Rubin podem ser descritas como se segue: em cada uma das **m** análises obtêm-se estimativas para um parâmetro de interesse Q, ou seja, Q_j para j=1, 2,...,m, podendo Q ser qualquer medida escalar a ser estimada, tal como média, correlação, coeficiente de regressão, etc.[9,16]. A estimativa geral será a média das estimativas individuais: $\overline{Q} = \frac{1}{m}\sum_{j=1}^{m}\hat{Q}_j$. Para a variância combinada, primeiramente calcula-se a variância média dentro das imputações: $\overline{U} = \frac{1}{m}\sum_{j=1}^{m}U_j$, onde U_j são as variâncias dos estimadores

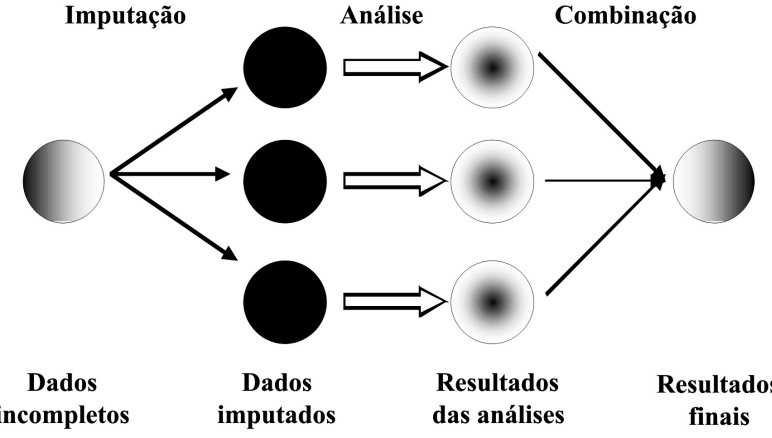

Figura 1 – Esquema da imputação múltipla (Figura extraída de www.multiple-imputation.com)

***Figure 1** – Representation of multiple imputation (Figure extracted from www.multiple-imputation.com)*

dentro de cada uma das imputações, para j=1, 2,...,m, e a variância entre imputações: $B = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{Q}_j - \bar{Q})^2$. Então a variância total, que é a variância combinada, será[6]: $T = \bar{U} + \left(1 + \frac{1}{m}\right)B$.

A imputação múltipla deste trabalho foi feita pelo método chamado *Bayesian Linear Regression* (*BLR*) (Pág.166-167, Rubin, 1987)[10]. Foi escolhido este método de imputação múltipla por ser adequado para a imputação de variáveis quantitativas. Sob o paradigma Bayesiano, este método parte do princípio de que as imputações múltiplas são feitas através de uma regressão linear múltipla ($Y=\alpha+\beta\mathbf{X}$), $Y \sim N(X\beta; I\sigma^2)$, em que a variável resposta Y será a variável a ser imputada, e resumidamente pode ser descrito como: os parâmetros β e σ a serem usados na imputação são estimados a partir de uma distribuição *a posteriori* própria. São calculados os valores preditos para os $y_{observados}$ e $y_{faltantes}$ e os valores usados para as imputações são os valores preditos para os $y_{faltantes}$ gerados pelas **m** repetições da estimação de β e σ.

A imputação múltipla pelo método *BLR* está implementada no pacote MICE[13] do programa R[18] que foi usado para este trabalho. Mais detalhes computacionais, tais como as instruções de programação, podem ser encontradas em Nunes et al.[4], onde é mostrado um estudo de simulação com a parte completa do banco de dados original.

Foi necessário um estudo detalhado sobre a suposição do mecanismo de não-resposta da variável albumina. Quando os dados são NMAR (não aleatório) é necessário que se incorpore explicitamente o mecanismo de não-resposta, algo que na maioria das situações se desconhece[19]. Já a suposição MAR (aleatório) tem um importante papel na questão de tratamento de dados faltantes, pois representa uma condição sob a qual inferências válidas podem ser obtidas sem se fazer referência ao mecanismo de não-resposta[7]. Infelizmente, é impossível determinar se os dados faltantes são MAR ou NMAR; isso pode ser simplesmente especulado e a suposição MAR não pode ser testada[7,8]. Assim, a análise de dados incompletos tem por característica depender de suposições que não podem ser testadas.

Para a escolha das variáveis que foram incluídas na imputação pelo método *BLR*, julgou-se adequado incluir variáveis que fizessem com que a suposição de que o mecanismo de não-resposta MAR fosse mais aceitável, pois, de acordo alguns autores, incluir tantos preditores quanto possível tende a tornar a suposição de MAR mais plausível, reduzindo assim a necessidade de se fazer ajustes especiais para mecanismos NMAR[17,20]. Neste caso, incluíram-se as variáveis ASA e o caráter da cirurgia nos modelos das IM, considerando-se a relação que apareceu entre essas variáveis e a não-resposta na variável albumina.

A Tabela 1 mostra como os dados faltantes da albumina estão distribuídos em relação as variáveis ASA e o caráter da cirurgia. No banco de dados, 122 pacientes não tinham a informação da albumina, ou seja,

Tabela 1 – Padrão dos dados faltantes de albumina (n=122) em relação a ASA e caráter de cirurgia. Percentual entre parênteses.
***Table 1** – Pattern of missing albumin data (n=122) in relation to ASA and characteristic of surgery. Percentage*

ASA	Caráter de cirurgia		Total n (%)
	Eletiva n (%)	Urgência n (%)	
ASA I/II	58 (47,5)	18 (14,8)	76 (62,3)
ASA III	17 (13,9)	15 (12,3)	32 (26,2)
ASA IV/V	3 (2,5)	11 (9,0)	14 (11,5)
Total	78 (63,9)	44 (36,1)	122 (100,0)

($\chi^2 = 17,64$; gl = 2; $p < 0,001$)

27,1% da amostra. Destes 122 pacientes, 63,9% eram pacientes que se submeteram à laparotomia eletiva, enquanto 36,1% sofreram cirurgia de urgência. Quanto à variável ASA, pode-se observar que quanto maior o nível ASA, menor o percentual de dados faltantes, variando de 62,3% a 11,5%. Em relação ao número total de observações faltantes (n = 122), 47,5% dos dados são de pacientes submetidos à cirurgia eletiva que se enquadraram na ASA I, enquanto somente 9,0% dos dados faltantes são de pacientes que fizeram cirurgia de urgência e pertenciam à ASA IV/V. Além disso, há associação entre à ASA e o caráter da cirurgia ($\chi^2 = 17,64$; gl = 2; $p < 0,001$). Ou seja, quanto pior o estado do paciente, menos dados faltantes se tem na variável albumina.

A suposição MAR feita neste trabalho baseou-se na ideia que aparece na Tabela 1, ou seja, que a ausência da informação da variável albumina pode estar relacionada com as variáveis ASA e o caráter da cirurgia, não importando o valor propriamente dito da albumina. Tal fato deixa menos provável a ideia de que se tenha um mecanismo de não-resposta NMAR, isto é, que a ausência da observação esteja relacionada com o valor da variável.

Foram feitas duas imputações múltiplas diferentes, uma não incluindo a variável desfecho do estudo original sobre escore de risco (óbito), e outra a incluindo, chamadas de IM(1) e IM(2), respectivamente. A ideia de se incluir o desfecho (óbito) é que, segundo Moons et al.[21], são preferíveis as imputações que incluem o desfecho no modelo. Considerando que a variável resposta é a albumina (Y_{imp}), ou seja, a variável a ser imputada, as imputações múltiplas podem ser descritas como:

IM(1): $(Y_{imp}) = \beta_1(\text{ASA III}) + \beta_2(\text{ASA IV/V}) + \beta_3(\text{Cirurgia urgente}) + \text{Constante}$

IM(2): $(Y_{imp}) = \beta_1(\text{ASA III}) + \beta_2(\text{ASA IV/V}) + \beta_3(\text{Cirurgia urgente}) + \beta_4(\text{Óbito=sim}) + \text{Constante}$.

Para a comparação dos métodos de imputação foram feitas regressões logísticas multivariáveis considerando-se como desfecho a variável óbito e como variáveis independentes as seguintes variáveis: ASA, tendo como categoria de referência ter ASA I ou II; a idade, sendo "até 75 anos" a categoria base; e albumina categorizada, sendo a categoria de referência 3,1g/dl. Depois de realizadas as imputações múltiplas, as estimativas gerais para os coeficientes β das regressões logísticas foram obtidas pela aplicação das Regras de Rubin citadas anteriormente. As comparações foram realizadas pela comparação dos valores das estimativas pontuais das razões de chances (RC), respectivos erros padrão e intervalos de confiança.

Todas as análises estatísticas foram realizadas no aplicativo R[18] versão 2.5.1.

Resultados

Na Tabela 2 observam-se os coeficientes estimados pelos modelos de regressão logística quando utilizadas diferentes estratégias de imputação. Para as variáveis/categorias ASA III, ASA IV/V e idade, os coeficientes resultaram bastante similares, independentemente da estratégia de imputação utilizada.

Especificamente para a albumina, houve uma diferença relevante que cabe ressaltar: os valores estimados pelos métodos de imputação múltipla tiveram valores bem próximos quando comparados entre si; entretanto, apresentaram valores inferiores aos estimados pelos métodos de imputação única. Enquanto pelos métodos de imputação única os valores foram 1,885 e 1,756 para a categoria "até 2,2 g/dl", respectivamente pelo método das medianas e do valor normal, o valor da IM(1) foi 1,501 e da IM(2) foi 1,611. Quando se observam os valores estimados para a categoria "2,3 a 3,0 g/dl" da albumina, nota-se que ocorre o mesmo que com a categoria anterior. Ou seja, para o método das medianas o valor foi 0,779 e pelo método do valor normal foi 0,773, enquanto para a IM(1) o valor foi 0,511 e para a IM(2) foi 0,553, conforme a Tabela 2.

Na Tabela 3 podem ser vistos os resultados das estimativas das razões de

Tabela 2 – Comparação entre os coeficientes do modelo de Regressão Logística obtidos com diferentes imputações dos valores faltantes da Albumina.
Table 2 *– Comparison among coefficients of the Logistic Regression model obtained with different imputations of the missing albumin values.*

Variável	Métodos de imputação			
	Medianas	Valor normal (3,5g/dl)	IM(1)	IM(2)
ASA III	1,201	1,201	1,267	1,269
ASA IV/V	3,105	3,136	3,231	3,212
Idade ≥ 75 anos	1,406	1,408	1,420	1,427
Alb até 2,2 g/dl	1,885	1,756	1,501	1,611
Alb 2,3 a 3,0 g/dl	0,779	0,773	0,511	0,553
Constante	-3,779	-3,670	-3,674	-3,728

chances, dos respectivos intervalos de 95% de confiança e erros padrões obtidos para o modelo de regressão logística múltipla, utilizando-se os diferentes métodos de imputação. Quando observados os valores obtidos para as categorias ASA III, ASA IV/V e para a variável idade, percebe-se que as quatro diferentes estratégias de imputação produziram estimativas bastante semelhantes para os parâmetros do modelo logístico ajustado, com exceção da razão de chances de ASA IV/V, que tanto para a IM(1) como IM(2) foram um pouco maiores que os valores obtidos nas imputações únicas. No entanto, a variabilidade foi semelhante em todas as estratégias de imputação utilizadas.

Para as estimativas das categorias de albumina, percebe-se que, quando o modelo de regressão logística ajustado levou em consideração as imputações múltiplas, os valores das razões de chances foram menores do que os obtidos pelos modelos que levaram em conta as imputações únicas. Vê-se que para a categoria "até 2,2 g/dl" da albumina, na IM(1), a razão de chances foi estimada em 4,5, com IC95% = [2,0;10,0], e para a IM(2) foi 5,0, com IC95% = [2,1;11,8], enquanto para as imputações únicas a mesma razão de chances ficou estimada em 6,6 com IC95% = [2,9;14,8] e 5,8 com IC95% = [2,8;11,8], respectivamente para o método das medianas e do valor normal. Ainda na Tabela 3 pode-se observar que os erros padrões para as categorias da variável albumina foram maiores nas estratégias das imputações múltiplas.

Tabela 3 – Estimativas da RC da regressão logística após as imputações.
Table 3 *– Estimates of OR of logistic regression after imputations.*

Variáveis independentes	RC [IC95%] e (Erro Padrão) dos modelos logísticos ajustados			
	Medianas	Valor normal (3,5)	IM(1)	IM(2)
ASA III	3,3[1,4;7,8] (0,438)	3,3[1,4;7,8] (0,438)	3,6[1,5;8,4] (0,439)	3,6[1,5;8,5] (0,441)
ASA IV/V	22,3[9,5;52,6] (0,438)	23,0[9,8;54,0] (0,435)	25,3[10,6;60,4] (0,444)	24,8[10,3;59,7] (0,448)
Idade ≥ 75 anos	4,1[1,8;9,1] (0,409)	4,1[1,8;9,1] (0,407)	4,1[1,9;9,2] (0,407)	4,2[1,9;9,3] (0,412)
Alb até 2,2 g/dl	6,6[2,9;14,8] (0,414)	5,8[2,8;11,8] (0,363)	4,5[2,0;10,0] (0,407)	5,0[2,1;11,8] (0,437)
Alb 2,3 a 3 g/dl	2,2[1,0;4,7] (0,389)	2,2[1,1;4,6] (0,357)	1,7[0,7;3,8] (0,417)	1,7[0,8;3,8] (0,403)

Com a ideia de investigar as diferenças constatadas entre as estimativas obtidas pelos quatro modelos logísticos que usaram dados com os diferentes métodos de imputação, foram feitas análises adicionais como as apresentadas a seguir na Figura 2, letras (a), (b) e (c) e na Tabela 3.

A Figura 2(a) mostra como ficou a distribuição da variável albumina categorizada em diferentes bancos de dados, isto é, com dados faltantes, com dados imputados pelas medianas, com dados imputados pelo limite inferior do valor normal (3,5 g/dl) e com as diferentes imputações múltiplas IM(1) e IM(2). É possível observar que a categoria "até 2,2 g/dl" teve menor frequência tanto na situação da imputação pelas medianas como para a imputação única do valor normal, quando comparada com o grupo sem imputação. Além disso, quando observada a categoria "2,3 a 3,0 g/dl", a imputação pelo valor normal também teve menor frequência que o grupo sem imputação. Isto ocorreu porque a imputação única pelas medianas só imputou valores para as categorias "2,3 a 3,0 g/dl" e "3,1 g/dl ou mais", pois os valores imputados foram 2,4 g/dl para o grupo de cirurgia urgente e 3,1 g/dl para o grupo de cirurgia eletiva e a imputação pelo valor normal só imputou o valor 3,5 g/dl.

Quando se observa na Figura 2(a) a distribuição das IM, percebe-se que houve valores imputados nas três categorias da albumina. Nas categorias extremas "até 2,2 g/dl" e "3,1 g/dl ou mais", a frequência diferiu daquela do grupo sem imputação; no entanto, as diferenças foram menores do que as referentes às imputações únicas. Para a categoria "2,3 a 3,0 g/dl", a distribuição das IM foi bastante próxima da situação "sem imputação". Salientando que existe a suposição de que os dados faltantes ocorreram ao acaso (MAR), então era esperado que isso acontecesse, ou seja, de que o padrão da albumina completada pela IM ficasse com uma distribuição parecida com a da albumina incompleta.

Para um maior detalhamento dos resultados, as Figuras 2(b) e 2(c) mostram a distribuição da variável albumina categorizada com os dados faltantes e com dados completados com as diferentes imputações em dois grupos: pacientes com cirurgia eletiva (Figura 2(b)) e pacientes com cirurgia de urgência (Figura 2(c)). Enquanto as IM tiveram distribuições bastante semelhantes à distribuição com dados faltantes (sem imputação), tanto na Figura 2(b) como na 2(c) as imputações únicas apresentaram discrepâncias maiores. Por exemplo, na

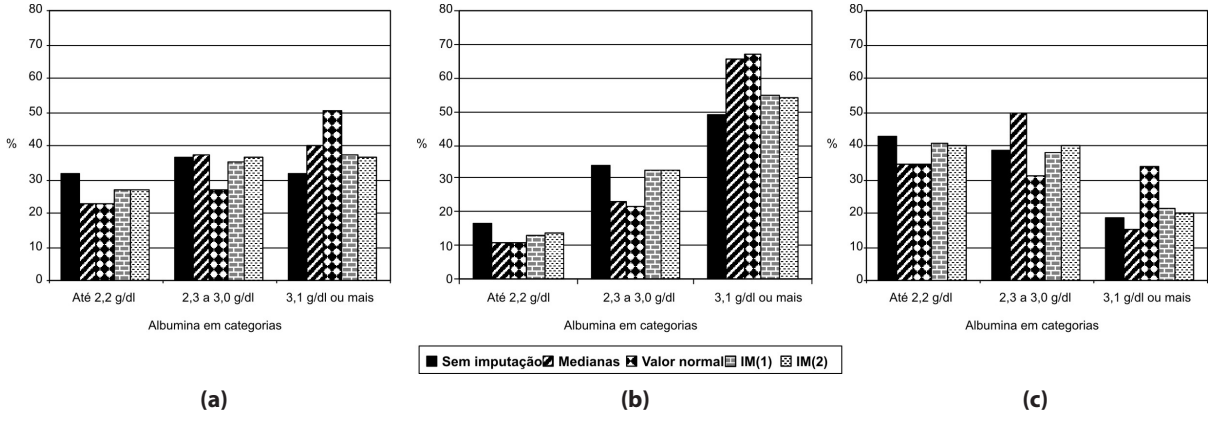

Figura 2 – Comparação das frequências relativas das categorias de albumina, considerando os dados sem imputação, imputados pelas medianas, imputados pelo limite inferior do valor normal (3,5 g/dl) e imputados pelas imputações múltiplas IM(1) e IM(2). Toda a amostra (a), grupo de cirurgia eletiva (b) e grupo de cirurgia urgente (c).

***Figure 2** – Comparison of relative frequencies of albumin categories, considering data without imputation, imputed by medians, imputed by lower limit of the normal value (3.5 g/dl) and imputed by multiple imputations IM(1) and IM(2). Entire sample (a), elective surgery group (b) and emergency surgery group (c).*

Figura 2(b), as frequências das imputações únicas para a categoria "até 2,2 g/dl" foram menores quando comparadas com a situação sem imputação, ocorrendo o oposto na categoria "3,1 g/dl ou mais".

Para o grupo de pacientes com cirurgia de urgência (Figura 2(c)) as maiores discrepâncias foram a frequência da imputação pela mediana na categoria "2,3 a 3,0 g/dl" e a frequência da imputação do valor normal na categoria "3,1 g/dl ou mais", ambas maiores em relação aos dados sem imputação.

Foram calculadas as medianas de acordo com o cruzamento das categorias de ASA e o caráter da cirurgia. Viu-se que os pacientes de ASA IV/V submetidos à cirurgia eletiva tiveram o valor mediano da albumina mais baixo (Md = 2,2 g/dl) que os das outras categorias de ASA (Md = 3,2 g/dl e Md = 3,1 g/dl para ASA I/II e III, respectivamente). Quanto ao caráter de cirurgia urgente, observa-se o mesmo padrão; entretanto, houve menor diferença entre os valores medianos, que variaram de 2,1 a 2,7 g/dl, sendo o menor valor o da categoria IV/V da ASA.

Discussão

Segundo Ambler et al.[11], muitos modelos de risco da área médica usam informações rotineiramente coletadas em hospitais que são referentes a preditores baseados nas características dos pacientes. Desfechos clínicos como morte hospitalar, por exemplo, em geral são informações completas. No entanto, muitos dos preditores têm observações faltantes. Frequentemente, os pacientes têm vários preditores sem informações, além de não ser incomum alguns preditores importantes terem mais que 50% de dados faltantes[22]. Esses problemas precisam ser levados em conta para que os modelos de risco possam ser confiáveis. Entretanto, a questão dos dados faltantes tem recebido pouca atenção dos pesquisadores[23].

Está claro que os dados faltantes podem afetar a predição dos modelos de risco e a opção de simplesmente ignorá-los e analisar somente os dados completos, resultando em tamanhos de amostras menores que o planejado, pode levar a viés nos resultados ou empobrecimento da predição, o que, na prática, afetaria as estratégias de tratamento e as decisões. No contexto de mortalidade cirúrgica, isso causaria sérias implicações clínicas[11]. Se, por causa dos dados faltantes, grupos inteiros de pacientes forem excluídos da análise, tais como grupos de idosos ou pacientes mais graves, ocorrerá um viés nos resultados. Portanto, tem sido recomendado que os dados faltantes sejam imputados antes de se criar os modelos de risco[5].

Os métodos de imputação única, que são regularmente usados na prática provavelmente devido à sua simplicidade, normalmente mostram um ganho em relação à análise restrita aos casos completos. Vale ressaltar que esse ganho pode depender da proporção de dados faltantes e do método de imputação única utilizado[5]. Entretanto, esses métodos podem reduzir a variabilidade amostral por imputarem valores do centro da distribuição (método das medianas) ou por imputarem um único valor (método do valor normal) para todos os pacientes com dados faltantes[24].

Apesar de, atualmente, a estratégia de imputação de dados já estar bastante difundida, a utilização da imputação múltipla ainda é muito incipiente, principalmente na área da saúde[1]. Talvez isso aconteça em decorrência da complexidade computacional que a imputação múltipla exige, principalmente quando comparada com os métodos de imputação única. Este trabalho apresentou resultados de imputações obtidas no R; entretanto, outros aplicativos estatísticos, tais como SPSS, SAS ou Stata, também têm implementado rotinas de imputação múltipla, o que indica a necessidade de se divulgar a técnica, pois já existe certa facilidade computacional para sua aplicação[4].

Uma vantagem da IM em relação à imputação única é o fato de a IM levar em conta a variabilidade entre imputações e, no caso do método *BLR*, por ter o componente Bayesiano embutido no procedimento, restringir a subestimação da variabilidade amostral, já que a cada vez (e são **m** vezes)

que é ajustada a regressão da IM, um valor diferente é gerado[11]. Segundo F. E. Harrel Jr.[5], quando há mais que 15% de perda de dados, na maior parte dos modelos é indicada a imputação múltipla. Portanto, como neste trabalho havia 27,1% de dados faltantes para a variável albumina, justifica-se a aplicação da IM.

Outra possível razão para a pouca utilização da IM pode ser a complexidade que envolve as suposições quanto aos mecanismos de não-resposta nos conjuntos de dados[6]. Infelizmente, são frequentes as situações onde há razões para se suspeitar que o mecanismo de não-resposta seja não aleatório (NMAR). Entretanto, dada a grande dificuldade em se incorporar nos modelos de imputação o mecanismo NMAR, é importante a realização de um estudo detalhado para se poder definir qual a melhor suposição para a não-resposta[19]. Quando é possível se supor aleatoriedade na não-resposta (mecanismo MAR), o processo de imputação fica facilitado, proporcionando ao pesquisador maior flexibilidade para a análise dos dados. Contudo, a suposição só pode ser feita por especulação, não havendo testes que comprovem sua violação ou não[8].

Verificou-se que o padrão de não-resposta da albumina tem a ver com as variáveis ASA e o caráter da cirurgia, onde, resumidamente, os pacientes mais graves têm menos dados faltantes na albumina. A partir dessa avaliação, considerou-se estas variáveis nos modelos das IM para tornar a suposição MAR mais factível. Por outro lado, seguindo a conclusão de Moons et al.[21], que constataram ser desejável a inclusão do desfecho no modelo de imputação, foi considerado também o desfecho óbito na IM(2). Entretanto, para o conjunto de dados estudado neste artigo, não houve uma diferença considerável nos resultados com a inclusão do desfecho[21].

Nas imputações múltiplas foram levados em conta a relação entre a albumina e as variáveis ASA e o caráter da cirurgia, o que não ocorreu nas imputações únicas. Isso provavelmente fez com que as estimativas dos modelos fossem diferentes. O método das medianas levou em conta o caráter da cirurgia, mas não considerou a variável ASA, sendo imputados valores que colocaram os pacientes em somente em duas categorias da albumina. O método do valor normal imputou somente um valor para a albumina, que foi 3,5 g/dl, ou seja, todos os pacientes com dados faltantes foram incluídos apenas em uma categoria da albumina. Quando observados os valores gerados pela imputação múltipla, percebe-se que os pacientes que tinham dados faltantes foram incluídos nas três categorias da variável albumina (vide Figura 2).

Essas diferenças nas imputações única e múltipla podem ter levado aos diferentes modelos obtidos, por exemplo, quando se observa a categoria "2,3 a 3,0 g/dl" da albumina; enquanto ela é significativa nos modelos com os dados imputados pelas imputações únicas, deixa de sê-lo quando o modelo logístico usa os dados imputados pelas IM. Ou seja, os modelos de risco para mortalidade cirúrgica seriam diferentes dependendo do método de imputação utilizado. Portanto, os pesquisadores devem investigar bem todas as possibilidades de tratamento dos dados faltantes[7]. Também é interessante ressaltar que, quanto à variabilidade das estimativas, os resultados foram consonantes com o que se encontra na literatura[2,9], ou seja, que as imputações únicas subestimam a variabilidade, gerando erros padrões menores que as imputações múltiplas.

Este trabalho mostrou a importância de se considerar métodos de imputação para dados faltantes, em especial a imputação múltipla. No entanto, deve-se ter cuidado na generalização dos resultados obtidos com este trabalho, já que eles foram obtidos em uma situação particular, usando o banco de dados como uma simples ferramenta para o exercício de se divulgar a técnica da imputação múltipla, em particular, o método *BLR*. Os tens relacionados ao tamanho amostral, tipo de variável e estrutura de relação entre as variáveis envolvidas devem ser sempre levados em consideração. Mais trabalhos são necessários para se avaliar o

desempenho de métodos de imputação, seja para conjuntos de dados com maior proporção de preditores contínuos, por exemplo, com possível relação não-linear, seja para conjuntos de dados com desfechos não-binários[11].

Referências

1. Zhou XH, Eckert GJ, Tierney WM. Multiple imputation in public health research. *Stat Med* 2001; 20(9-10): 1541-9.

2. Rubin DB. Multiple imputation after 18+ years. *JASA* 1996; 91: 473-89.

3. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychol Methods* 2002; 7: 147-77.

4. Nunes LN, Klück MM, Fachel JMG. Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. *Cad Saúde Pública* 2009; 25: 268-78.

5. Harrell FE Jr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, And Survival Analysis.* Springer–Verlag: New York; 2001.

6. Tang L, Song J, Belin TR, Unutzer J. A comparison of imputation methods in a longitudinal randomized clinical trial. *Stat Med* 2005; 24: 2111-28.

7. Kenward MG e Carpenter J. Multiple imputation: current perspectives. *Stat Methods Med Res* 2007; 16(3): 199-218.

8. Van der Heijden GJ, Donders AR, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *J Clin Epidemiol* 2006; 59(10): 1102-9.

9. Schafer JL. *Analysis of incomplete multivariate data.* London: Chapman & Hall; 1997.

10. Rubin DB. *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley; 1987.

11. Ambler G, Omar RZ, Royston P. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Stat Methods Med Res* 2007; 16(3): 277-98.

12. White IA, Wood A, Royston P. Editorial: Multiple imputation in practice. *Stat Methods Med Res* 2007; 16; 195-7.

13. Klück M. *Metodologia para ajuste de indicadores de desfechos hospitalares por risco prévio do paciente* [tese de doutorado]. Faculdade de Medicina da Universidade Federal do Rio Grande do Sul: Porto Alegre; 2004.

14. Little RJA. Regression with Missing Xs - A Review. *JASA* 1992; 87(420): 227-37.

15. Zhang P. Multiple imputation: Theory and method. *Int Stat Rev* 2003; 71(3): 581-92.

16. Bernaards AB, Farmer MM, Qi K, Dulai GS, Ganz PA, Kahn KL. Comparison of two multiple imputation procedures in a cancer screening survey. *J Data Sci* 2003; 1: 293-312.

17. Van Buuren S, Oudshoorn CGM. *Multivariate imputation by chained equations. MICE V1.0 User's Manual*. Report PG/VGZ/00.038. Leiden: TNO Preventie en Gezondheid; 2000.

18. R Development Core Team R. A language and environment for statistical computing. Vienna, Austria: R *Foundation for Statistical Computing.* 2004. Disponível em http://www.R-project.org ISBN 3-900051-00-3. [Acessado em 2 de setembro de 2007]

19. Allison, PD. *Missing Data*. Thousand Oaks, CA: Sage; 2001.

20. Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med* 1999; 18: 681-94.

21. Moons KG, Donders RA, Stijnen T, Harrell FE Jr. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol* 2006; 59(10): 1092-101.

22. Ambler G, Omar RZ, Royston P, Kinsman R, Keogh B, Taylor KM. A generic, simple risk stratification model for heart valve surgery. *Circulation* 2005; 112: 224-31.

23. Clark, TG and Altman, DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. *J Clinic Epidemiol* 2003; 56 (1): 28-37.

24. Little, RJA e Rubin, DB. *Statistical analysis with missing data.* 2nd ed. New York: Wiley; 2002.

Recebido em: 06/04/09
Versão final reapresentada em: 13/07/10
Aprovado em: 14/07/10