Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos

Multiple imputations for missing data: a simulation with epidemiological data

Luciana Neves Nunes [1,2]
Mariza Machado Klück [1]
Jandyra Maria Guimarães Fachel [1,2]

[1] Faculdade de Medicina, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil.
[2] Instituto de Matemática, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil.

Correspondência
L. N. Nunes
Departamento de Estatística, Instituto de Matemática, Universidade Federal do Rio Grande do Sul.
Av. Bento Gonçalves 9500, Porto Alegre, RS 91509-900, Brasil.
lununes@mat.ufrgs.br

Abstract

In situations with missing data, statistical analyses are usually limited to subjects with complete data. However, such estimates may be biased. The method of "filling in" missing data is called imputation. This article aimed to present a multiple imputation method. From a data set of 470 surgical patients, logistic models were developed for death as the outcome. Two incomplete data sets were generated: one with 5% and another with 20% of missing data in a single variable. Logistic models were fitted for the complete and incomplete data sets and for the data set completed by multiple imputations. Estimates obtained for the data set with missing data were different from those observed in the complete data set, mainly in the situation with 20% of missing data. The multiple imputation used here appeared efficient, producing very similar results to those obtained with the complete data set. However, one coefficient became non-significant. The analysis using multiple imputations was considered superior to using the data sets that excluded incomplete cases from the analysis.

Statistical Data Interpretation; Statistical Models; Database

Introdução

Um problema comum em investigações científicas é a ocorrência de dados faltantes (*missing data*), especialmente na área da Saúde e das Ciências Sociais [1]. Determinar a abordagem analítica adequada para conjuntos de dados com observações incompletas é uma questão que pode ser bastante delicada, pois a utilização de métodos inadequados pode levar a conclusões erradas sobre o fenômeno na população. O desenvolvimento de métodos estatísticos direcionados a solucionar problemas de dados faltantes tem sido uma área de pesquisa bastante ativa nas últimas décadas [2,3,4,5].

A perda de dados é um grande desafio no planejamento e análise dos estudos epidemiológicos, nos quais, freqüentemente, o objetivo é determinar preditores que contribuem para prever a ausência ou presença de uma doença em uma população. Perda de informações, tanto nos preditores como no desfecho, pode levar a problemas sérios na análise dos dados. Portanto, é importante que se estabeleçam estratégias para lidar com dados faltantes, seja planejando a pesquisa com o máximo de esforço para evitar perda de informações, seja abordando os dados faltantes com técnicas adequadas desenvolvidas para contornar este problema [6].

É comum que se encontrem diferenças nos modelos obtidos com a análise de casos completos – abordagem muito comum em estudos

epidemiológicos – em relação aos modelos obtidos com os dados imputados no que se refere aos preditores selecionados, coeficientes de regressão e correspondentes erros padrão [6].

Para contornar esse problema, desde os anos 80 surgiram técnicas estatísticas que envolvem imputação de dados faltantes. Essas técnicas têm por objetivo "completar" os dados faltantes e possibilitar a análise com todos os indivíduos do estudo. As primeiras técnicas de imputação desenvolvidas envolviam métodos relativamente simples, tais como, substituição dos dados faltantes pela média, pela mediana, por interpolação ou até por regressão linear. Todas essas técnicas mencionadas permitem "preencher" os dados faltantes por meio do que se chama de imputação única, ou seja, o dado ausente é preenchido uma única vez e então se utiliza o banco de dados completo para as análises. Entretanto, a incerteza associada à imputação deve ser levada em conta para que os resultados obtidos com os dados completos sejam válidos, pois os valores imputados não são valores reais. Para solucionar essa questão foi desenvolvida a técnica de Imputação Múltipla (IM) [2].

A literatura sobre imputação múltipla tem se expandido muito desde o início da década de 90 [7]. No PubMed, uma busca com a palavra-chave "imputation" indicou 571 trabalhos publicados (1º de setembro de 2007), sendo que só nos últimos 12 meses foram 66 trabalhos. Ainda, importantes periódicos perceberam a necessidade de fazer edições especialmente dedicadas ao assunto de imputação. Por exemplo, as revistas *Statistics in Medicine*, em 1997, *Statistica Neerlandica*, em 2003, *Journal of Clinical Epidemiology*, em 2006, e mais recentemente, *Statistical Methods in Medical Research*, em junho de 2007, dedicaram edições exclusivas a artigos sobre tratamentos a dados faltantes e principalmente à IM. Esses fatos revelam que o estudo de metodologias para dados faltantes vem sendo bastante debatido mais recentemente, o que indica a pertinência deste trabalho.

A IM está se tornando o método cada vez mais popular para tratar dados faltantes. Isso se deve principalmente à sua enorme flexibilidade – se bem usada, pode lidar com dados faltantes de todos os tipos (quantitativos, categóricos ordinais, nominais etc.). Também é válida para dados desempenhando diferentes papéis nos modelos (preditores, confundimento, desfecho etc.). Por separar a tarefa de análise em duas etapas (imputação e análise dos dados completos), a sua utilização é simplificada [7].

Desde a sua introdução há mais ou menos trinta anos, a IM se tornou uma abordagem importante e influente na análise de dados incompletos. Durante esse período, o número de aplicações tem crescido, incluindo a análise de estudos observacionais na área de saúde pública e ensaios clínicos. Em paralelo a esse desenvolvimento, ferramentas de IM têm sido incorporadas em muitos aplicativos estatísticos. Inevitavelmente, seu crescente uso tem gerado novas discussões e desafios [8].

A proposta deste trabalho é promover uma maior divulgação do método de IM para os pesquisadores da área da saúde e também mostrar que o pesquisador tem um ganho considerável em suas análises quando decide imputar os dados faltantes em vez de fazer a análise restrita aos casos completos. A comparação dos resultados com dados imputados por imputação múltipla e da análise com casos completos será feita usando-se um conjunto de dados reais.

Método

Imputação múltipla

D. B. Rubin, ainda nos anos 70, propôs a técnica de IM para resolver o problema de não-resposta em pesquisas. No entanto, apenas recentemente essa técnica vem sendo mais utilizada devido aos desenvolvimentos computacionais para sua implementação. A técnica possibilita a inclusão da incerteza da imputação nos resultados, corrigindo o maior problema associado à imputação única [2]. A IM consiste em três passos:

1. São obtidos *m* bancos de dados completos por meio de técnicas adequadas de imputação;
2. Separadamente, os *m* bancos são analisados por um método estatístico tradicional, como se realmente fossem conjuntos completos de dados;
3. Os *m* resultados encontrados no passo 2 são combinados de um jeito simples e apropriado para obter a chamada inferência da imputação repetida.

O primeiro passo é a parte fundamental da IM, pois as técnicas de imputação utilizadas têm de preservar a relação das observações faltantes e presentes e ainda levar em conta o mecanismo de ausência e o padrão dos dados faltantes. Os mecanismos dividem-se 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); e os padrões são: monotônicos e não-monotônicos [2,9].

A partir das *m* imputações realizadas, o passo 2 da IM pode ser realizado, ou seja, os *m* bancos de dados são analisados por métodos tradicionais de análise. Finalmente, os *m* resultados obtidos podem ser combinados usando-se as regras propostas por Rubin [2].

As regras de Rubin (*Rubin rules*) estão amplamente divulgadas na literatura que trata de IM, pois são normas simples que resolvem o passo 3 da IM. Essas regras podem ser usadas independentemente do método utilizado para fazer a IM [9]. A idéia é que a partir de cada análise sejam obtidas as estimativas para o parâmetro de interesse Q, ou seja, Q_j para j = 1, 2,...,m. Segundo Schafer [4], Q pode ser qualquer medida escalar a ser estimada, tal como média, correlação, coeficiente de regressão ou razão de chances. Então, a estimativa combinada 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 dentro das imputações: $\overline{U} = \frac{1}{m}\sum_{j=1}^{m} U_j$ e a variância entre imputações: $B = \frac{1}{m-1}\sum_{j=1}^{m} \left(\hat{Q}_j - \overline{Q}\right)^2$. Então, a variância total, que é a variância combinada, será: $T = \overline{U} + \left(1+\frac{1}{m}\right) B$.

Para a realização da análise computacional, alguns aplicativos têm sido bastante citados na literatura, pois disponibilizam o uso dos métodos de IM. Dentre os mais utilizados, pode-se citar o SAS, S-Plus, SOLAS, NORM, BMDP e MICE, sendo que o MICE é de domínio público, pois é operado dentro do ambiente do aplicativo R (The R Foundation for Statistical Computing, Viena, Áustria; http://www.r-project.org). Análises do desempenho dos aplicativos computacionais para IM têm sido publicadas na literatura [10,11,12].

Segundo Harrell Jr. [13], é possível definir linhas gerais para a escolha entre os métodos de imputação de acordo com a proporção de dados faltantes em alguma das variáveis:

a) Proporção ≤ 0,05 → Neste caso pode ser usada a imputação única ou analisar somente os dados completos;

b) Proporção entre 0,05 e 0,15 → Imputação única pode ser usada provavelmente sem problemas, entretanto o uso da imputação múltipla é indicado;

c) Proporção ≥ 0,15 → A imputação múltipla é indicada na maior parte dos modelos.

No caso de haver muitos preditores com dados faltantes, devem ser feitas as mesmas considerações acima, mas os efeitos das imputações serão mais pronunciados.

Fonte de dados

Este artigo utilizará como exemplo um banco de dados cedido por Klück [14], trabalho em que foi desenvolvido e validado um escore de risco multifatorial para mortalidade cirúrgica pós-laparotomia exploradora. A população de pesquisa foi composta por pacientes internados no Hospital de Clínicas de Porto Alegre (HCPA), Porto Alegre, Rio Grande do Sul, de fevereiro de 2000 a dezembro de 2003. O desfecho estudado foi óbito num período de até 30 dias após a realização da cirurgia. O banco de dados original é composto por 651 pacientes. Entretanto, para ilustrar os métodos de IM neste artigo, foram utilizados somente 470 pacientes que tinham todas as variáveis de interesse completas.

Pelo fato do desfecho ser binário (óbito/não-óbito), o modelo utilizado foi o de regressão logística. A partir da coorte de derivação, foi feita a modelagem conforme descrito por Klück [14]. O modelo final obtido incluiu as variáveis: idade (< 75 anos e ≥ 75 anos), albumina sérica em três categorias (até 2,2; 2,3 a 3,0 e > 3,0g/dl) e ASA em três grupos (ASA I/II, ASA III e ASA IV/V). A classificação ASA (American Society of Anestesiology) é uma avaliação pré-anestésica e segue o seguinte: ASA I: paciente saudável, sem doença sistêmica e fora dos extremos de idade; ASA II: indivíduo com uma doença sistêmica bem controlada, que não afeta a sua atividade diária, ou paciente com um risco anestésico como tabagismo, obesidade ou alcoolismo; ASA III: indivíduo com múltiplas doenças sistêmicas ou com uma doença sistêmica grave, que limite a sua atividade diária; ASA IV: indivíduo com doença severa e incapacitante, em estágio terminal, ou mal controlada; e ASA V: paciente em iminente risco de morte, sendo a cirurgia o último recurso possível para preservar a vida ou atenuar o sofrimento.

A opção de usar somente os pacientes com informações completas foi feita para que fosse possível comparar os resultados do banco completo com os resultados dos bancos imputados e, assim, poder avaliar o método da imputação múltipla. Dessa maneira, os resultados obtidos com a análise do banco completo foram considerados valores verdadeiros.

Com base nesse banco completo, foram criados, por simulação, dois bancos de dados incompletos em que se excluiu, aleatoriamente, cerca de 5% e 20% das observações da variável albumina pelo gerador aleatório do SPSS 13.0 (SPSS Inc., Chicago, Estados Unidos). Por não se utilizar nenhum critério *a priori* para a exclusão das observações, pode-se dizer que o mecanismo que gerou esses dados faltantes foi MCAR. Nesse caso, o padrão da não-resposta é monotônico.

A variável albumina foi escolhida para ter observações excluídas porque foi uma variável que originalmente teve dados faltantes. O banco de dados com 5% de observações faltantes será referido como Banco Incompleto 5 (BI-5), enquanto que o banco de dados com 20% de observações

faltantes será referido como Banco Incompleto 20 (BI-20).

O modelo de IM é ajustado sob o paradigma Bayesiano, isto é, a partir do resultado da distribuição *a posteriori*, um conjunto de extrações aleatórias é feito para as observações faltantes a partir dos dados observados, obtendo-se assim o banco completo. Esse processo é repetido m vezes, resultando m bancos completos. Neste trabalho, foram considerados dois métodos de IM que partem do mesmo princípio, ou seja, de que as imputações múltiplas são feitas a partir de uma regressão linear ($Y = \alpha+\beta X$), $Y \sim N(X\beta; I\sigma^2)$, em que a variável-resposta Y será a variável a ser imputada. Resumidamente, as imputações são realizadas seguindo os métodos descritos a seguir, conforme Rubin [2]:

- *Preditive Mean Matching* (PMM)[2]: os parâmetros são estimados a partir de uma distribuição *a posteriori* própria. São calculados os valores preditos para os $y_{observados}$ e $y_{faltantes}$. Para cada $y_{faltante}$ predito, procura-se a unidade observada com valor predito mais próximo, e utiliza-se o valor observado como valor a ser imputado. A variabilidade entre imputações é gerada por meio dos passos que servem para estimar β e σ e que são repetidos m vezes;
- *Bayesian Linear Regression* (*BLR*) [2]: assim como o método PMM, são estimados β e σ, mas os m valores usados para as imputações são os próprios valores preditos para os $y_{faltantes}$ gerados por m repetições da estimação de β e σ.

Esses métodos foram escolhidos por serem adequados para a imputação de variáveis quantitativas, como no caso da variável albumina. Para a análise, foram feitos três diferentes modelos de regressão, IM(1), IM(2) e IM(3), tendo como variável-resposta (Y_{imp}) a albumina, com o objetivo de comparar os resultados obtidos, sendo que para cada uma das regressões foram feitas as imputações pelo método PMM e BLR. Os modelos foram:

IM(1): Albumina = β_1(ASA III)+β_2(ASA IV/V)+β_3(Idade ≥75)+constante

IM(2): Albumina = β_1(ASA III)+β_2(ASA IV/V)+β_3(Cirurgia urgente)+constante

IM(3): Albumina = β_1(ASA III)+β_2(ASA IV/V)+β_3(Idade ≥75)+β_4(Cirurgia urgente)+constante

A IM foi realizada usando-se o pacote Multivariate Imputation by Chained Equatoins (MICE) [15] do programa R. Maiores detalhes computacionais podem ser vistos no subitem *Algoritmo* deste artigo.

Quanto à inclusão de variáveis no modelo de imputação, van Buuren et al. [16] sugerem, como regra geral, usar toda a informação das variáveis disponíveis, o que produz imputações múltiplas com mínimo viés e máxima precisão. Esse princípio implica que o número de preditores seja tão grande quanto possível. Alguns autores observam que incluir tantos preditores quantos forem possíveis tende a fazer a suposição de MAR mais plausível, reduzindo a necessidade de se fazer ajustes especiais para mecanismos NMAR [15,16].

Segundo van Buuren et al. [16], pode-se usar a seguinte estratégia de seleção de co-variáveis:

1) Incluir todas as co-variáveis que aparecem no modelo de dados completos;
2) Adicionalmente, incluir fatores que influenciaram a ocorrência dos dados faltantes. Outras variáveis de interesse são aquelas para as quais as distribuições diferem entre os grupos de respondentes e não-respondentes. Essas podem ser encontradas inspecionando-se suas correlações/associações com a resposta indicadora da variável-alvo (isto é, a variável a ser imputada);
3) Remover as variáveis selecionadas no passo 2 que apresentem muitos dados faltantes dentro do subgrupo de casos incompletos. Um indicador simples é o percentual de casos observados dentro desse subgrupo.

Usualmente, muitos preditores utilizados para imputação são eles mesmos incompletos. A princípio, poderia ser aplicada a modelagem acima para cada preditor incompleto, mas isto poderia levar a problemas em cascata. Na prática, freqüentemente existe um pequeno conjunto de variáveis-chave, para as quais a imputação é necessária, o que sugere que os passos 1 a 3 devem ser feitos somente para estas variáveis-chave.

Após a realização da IM com m igual a 5 imputações, opção *default* do MICE [15], os bancos de dados completados foram analisados por regressão logística, sendo o desfecho óbito sim ou não e tendo como variáveis independentes: ASA (ASA I/II – categoria de referência, ASA III e ASA IV/V), idade (< 75 anos – categoria de referência e ≥ 75 anos) e albumina (≤ 2,2; 2,3 a 3 e ≥ 3,1g/dl – categoria de referência), seguindo o modelo obtido por Klück [14]. As estimativas gerais foram obtidas pela aplicação das Regras de Rubin citadas anteriormente e implementadas em uma planilha do Excel (Microsoft Corp., Estados Unidos).

Algoritmo

Para a realização das imputações expostas acima, as instruções para o uso do R foram as seguintes:

```
library(mice)
imputacao <- read.spss("C:/Lu/Tese/alb_imp80_
reg1.sav",
to.data.frame=T,use.value.labels=F)
summary(imputacao)
md.pattern(imputacao)
imp <- mice(imputacao)
```

```
imp
imp$imp$ALBUMINA
```

O programa R foi desenvolvido no Projeto R (*R Project*). É um programa gratuito e de código aberto, e a página oficial do projeto está em: http://www.r-project.org. Há também um espelho (*mirror*) brasileiro da área de *downloads* do programa no Departamento de Estatística da Universidade Federal do Paraná: http://www.est.ufpr.br/R. Os comandos apresentados mostram como foi feita a imputação múltipla pelo método PMM para o BI-20. Para a imputação múltipla pelo método BLR, somente o comando imp sofre modificação, ficando: imp<-mice(imputacao, defaultImp=c("norm")).

Resultados

A Tabela 1 mostra os resultados conseguidos com o banco de dados completo e com os bancos incompletos, obtidos por simulação com 5% e com 20% de perda, isto é, antes de serem completados pelos métodos de imputação (BI-5 e BI-20). É possível observar que, com exceção da categoria "até 2,2g/dl" da variável albumina, as razões de chances (RC) do banco completo foram levemente maiores que as do BI-5. Também na Tabela 1 nota-se que, com exceção da categoria III da variável ASA, os valores das RC obtidos com o BI-20 foram superestimados quando comparados com o banco real completo, e que seus respectivos intervalos de confiança (IC) foram notadamente mais amplos. Quanto aos modelos logísticos ajustados, percebe-se que as variáveis incluídas foram significativas, com exceção da categoria "2,3 a 3,0g/dl" da albumina, que tem como limite inferior o valor um quando ajustado o modelo com o BI-5.

Na Tabela 2 são apresentados os resultados da regressão logística utilizando-se os valores imputados pelo método de IM para a albumina, para o banco incompleto com 5% de dados faltantes (BI-5), usando-se diferentes configurações do modelo de regressão a ser utilizado pelo método PMM. Observa-se que os valores das estimativas ficaram bastante próximos daqueles estimados pelo banco completo (n = 470) para quase todas as variáveis e categorias, inclusive para ASA IV/V, para a qual houve 20% na RC quando observada a Tabela 1. As amplitudes dos intervalos de confiança são também praticamente equivalentes.

A Tabela 3 mostra os resultados obtidos para a regressão logística com a variável albumina imputada por meio do método BLR, com diferentes configurações do modelo de regressão linear para obtenção do valor predito, para o banco de dados BI-5. É possível observar que os valores estimados pelos bancos com dados imputados tiveram bastante similaridade com os valores estimados pelo banco completo. As estimativas pontuais das RC que usaram os dados imputados pela IM(1) foram exatamente iguais para as variáveis ASA III, ASA IV/V e idade, apresentando os valores 3,4; 20,2 e 2,9, respectivamente, e os ICs foram iguais para as variáveis ASA III e idade. Os valores dos erros-padrão foram levemente maiores para todas as variáveis quando se usou imputação, se comparados com os valores estimados pelo banco completo. Com exceção da categoria "2,3 a 3 g/dl" da albumina, todas as demais variáveis foram significativas no modelo logístico com dados imputados.

Os resultados apresentados na Tabela 4 são os obtidos com a variável albumina imputada pelo método PMM no BI-20. Os ICs para as RC foram, em sua maioria, mais largos que os estimados pelo banco completo. Cabe ressaltar que a categoria "2,3 a 3,0g/dl" da variável albumina deixou de ser significativa em todos os modelos com dados imputados.

Quando observadas as Tabelas 2 e 3, que consideram os resultados da regressão logística para dados de albumina imputados pelo método de IM para o banco com simulação de 5% dos dados faltantes, pode-se fazer uma comparação dos resultados obtidos pelos dois métodos de IM utilizados neste trabalho (PMM e BLR). As estimativas pontuais das RC e os respectivos ICs foram, em geral, muito parecidos para os dois métodos de IM. Quando observada a RC estimada pela IM(3) no método BLR para a categoria ASA III, percebe-se que o valor RC = 19,6 foi um pouco discrepante em relação a todos os valores estimados pelo método PMM (RC = 20,4 para IM(1), RC = 20,1 para IM(2) e RC = 20,4 para IM(3)) e mesmo em relação aos coeficientes estimados pelas IM (1) (RC = 20,2) e IM (2) (RC = 20,3) do método BLR.

A Tabela 5 apresenta os resultados das regressões logísticas ajustadas com a albumina imputada pelo método BLR para o banco em que foram simuladas 20% das perdas BI-20. Quando se observa as estimativas das RC e seus respectivos ICs, percebe-se que os valores estimados foram bastante semelhantes. Os erros-padrão, quando comparados com as estimativas do banco completo, apresentaram-se maiores para todas as variáveis. Novamente, a categoria "2,3 a 3g/dl" da albumina deixou de ser significativa em todas os modelos que usaram dados imputados.

Ao se comparar as Tabelas 4 e 5, que consideram os resultados obtidos nos modelos de regressão logística com 20% dos dados faltantes de albumina imputados, é possível comparar os resultados conseguidos pelos dois métodos de IM

Tabela 1

Estimativas da regressão logística para o banco de dados completo e bancos incompletos (BI-5 e BI-20). Modelos ajustados com as mesmas variáveis independentes e desfecho óbito.

Variáveis independentes	RC [IC95%] e (erro padrão) dos modelos logísticos ajustados		
	Banco completo (n = 470)	BI-5 (n = 440)	BI-20 (n = 383)
ASA III	3,4 [1,5; 7,9]	3,0 [1,3; 6,2]	3,3 [1,3; 8,7]
	-0,422	-0,428	-0,489
ASA IV/V	20,2 [8,8; 46,0]	16,4 [7,1; 37,8]	22,3 [8,6; 57,6]
	-0,421	-0,426	-0,484
Idade ≥ 75	2,9 [1,5; 5,8]	2,7 [1,3; 5,6]	4,0 [1,8; 8,9]
	-0,348	-0,363	-0,405
Albumina até 2,2g/dl	5,3 [2,7; 10,5]	5,6 [2,7; 11,7]	7,0 [3,1; 15,9]
	-0,349	-0,368	-0,42
Albumina 2,3 a 3,0g/dl	2,1 [1,1; 4,1]	2,0 [1,0; 4,2]	2,6 [1,2; 5,9]
	-0,345	-0,364	-0,412

Tabela 2

Estimativas da regressão logística após imputações múltiplas pelo método PMM em diferentes regressões. Mecanismo MCAR (BI-5; n = 440).

Variáveis independentes	RC [IC95%] e (erro padrão) dos modelos logísticos ajustados			
	Banco completo	IM(1) *	IM(2) **	IM(3) ***
ASA III	3,4 [1,5; 7,9]	3,6 [1,5; 8,1]	3,5 [1,5; 8,1]	3,5 [1,5; 8,1]
	-0,422	-0,423	-0,423	-0,423
ASA IV/V	20,2 [8,8; 46,0]	20,4 [8,8; 47,0]	20,1 [8,8; 46,1]	20,4 [8,9; 46,8]
	-0,421	-0,427	-0,423	-0,424
Idade ≥ 75	2,9 [1,5; 5,8]	2,9 [1,5; 5,9]	2,9 [1,4; 5,7]	2,9 [1,5; 5,8]
	-0,348	-0,352	-0,351	-0,348
Albumina até 2,2g/dl	5,3 [2,7; 10,5]	5,1 [2,4; 10,8]	5,2 [2,6; 10,6]	5,1 [2,5; 10,2]
	-0,349	-0,378	-0,359	-0,357
Albumina 2,3 a 3,0g/dl	2,1 [1,1; 4,1]	1,9 [0,9; 3,9]	1,9 [0,9; 4,0]	1,9 [0,9; 3,8]
	-0,345	-0,375	-0,367	-0,36

* IM(1): Albumina = β_1(ASA III)+β_2(ASA IV/V)+β_3(Idade ≥75)+cte;

** IM(2): Albumina = β_1(ASA III)+β_2(ASA IV/V)+β_3(Cirurgia urgente)+cte;

*** IM(3): Albumina = β_1(ASA III)+β_2(ASA IV/V)+β_3(Idade ≥75)+β^4(Cirurgia urgente)+cte.

utilizados (PMM e BLR). O modelo obtido com a IM(1) pelo método BLR foi um pouco diferente dos demais, tanto os ajustados com os dados imputados pelo método PMM como os modelos das IM(2) e IM(3) do método BLR, que foram razoavelmente semelhantes entre si. Na IM(1) do BLR, os valores das RC das categorias ASA III e ASA IV/V, e da categoria "até 2,2 g/dl" foram menores e maior, respectivamente, em relação aos valores das RC estimadas pelos outros modelos que usam dados imputados.

Discussão

Vale ressaltar que, sendo o objetivo principal deste artigo divulgar métodos de imputação múltipla, não serão discutidos em detalhe os resultados epidemiológicos e suas implicações, mas somente os aspectos estatísticos da análise.

Quando comparados os resultados obtidos com o banco de dados original (sem dados faltantes) com os resultados da análise feita nos bancos nos quais foram simuladas faltas de da-

Tabela 3

Estimativas da regressão logística após imputações múltiplas pelo método BLR em diferentes regressões. Mecanismo MCAR (BI-5, n = 440).

Variáveis independentes	RC [IC95%] e (erro padrão) dos modelos logísticos ajustados			
	Banco completo	IM(1) *	IM(2) **	IM(3) ***
ASA III	3,4 [1,5; 7,9]	3,4 [1,5; 7,9]	3,5 [1,5; 8,1]	3,4 [1,5; 7,9]
	-0,422	-0,422	-0,423	-0,424
ASA IV/V	20,2 [8,8; 46,0]	20,2 [8,8; 46,1]	20,3 [8,9; 46,6]	19,6 [8,5; 45,2]
	-0,421	-0,423	-0,423	-0,425
Idade ≥ 75	2,9 [1,5; 5,8]	2,9 [1,5; 5,8]	2,9 [1,5; 5,8]	2,9 [1,5; 5,8]
	-0,348	-0,35	-0,351	-0,351
Albumina até 2,2g/dl	5,3 [2,7; 10,5]	5,2 [2,5; 10,4]	5,1 [2,5; 10,3]	5,5 [2,7; 11,2]
	-0,349	-0,36	-0,362	-0,363
Albumina 2,3 a 3,0g/dl	2,1 [1,1; 4,1]	2,0 [1,0; 4,1]	1,9 [0,9; 3,9]	2,0 [1,0; 4,1]
	-0,345	-0,359	-0,361	-0,363

* IM(1): Albumina = β_1(ASA III)+β_2(ASA IV/V)+β_3(Idade ≥75)+cte;

** IM(2): Albumina = β_1(ASA III)+β_2(ASA IV/V)+β_3(Cirurgia urgente)+cte;

*** IM(3): Albumina = β_1(ASA III)+β_2(ASA IV/V)+β_3(Idade ≥75)+β^4(Cirurgia urgente)+cte.

Tabela 4

Estimativas da regressão logística após imputações múltiplas pelo método PMM em diferentes regressões. Mecanismo MCAR (BI-20, n = 383).

Variáveis independentes	RC [IC95%] e (erro padrão) dos modelos logísticos ajustados			
	Banco completo	IM(1) *	IM(2) **	IM(3) ***
ASA III	3,4 [1,5; 7,9]	3,4 [1,5; 7,9]	3,5 [1,5; 8,1]	3,4 [1,5; 7,9]
	-0,422	-0,422	-0,423	-0,424
ASA IV/V	20,2 [8,8; 46,0]	20,2 [8,8; 46,1]	20,3 [8,9; 46,6]	19,6 [8,5; 45,2]
	-0,421	-0,423	-0,423	-0,425
Idade ≥ 75	2,9 [1,5; 5,8]	2,9 [1,5; 5,8]	2,9 [1,5; 5,8]	2,9 [1,5; 5,8]
	-0,348	-0,35	-0,351	-0,351
Albumina até 2,2g/dl	5,3 [2,7; 10,5]	5,2 [2,5; 10,4]	5,1 [2,5; 10,3]	5,5 [2,7; 11,2]
	-0,349	-0,36	-0,362	-0,363
Albumina 2,3 a 3,0g/dl	2,1 [1,1; 4,1]	2,0 [1,0; 4,1]	1,9 [0,9; 3,9]	2,0 [1,0; 4,1]
	-0,345	-0,359	-0,361	-0,363

* IM(1): Albumina = β_1(ASA III)+β_2(ASA IV/V)+β_3(Idade ≥75)+cte;

** IM(2): Albumina = β_1(ASA III)+β_2(ASA IV/V)+β_3(Cirurgia urgente)+cte;

*** IM(3): Albumina = β_1(ASA III)+β_2(ASA IV/V)+β_3(Idade ≥75)+β^4(Cirurgia urgente)+cte.

dos e estes indivíduos retirados, houve discrepâncias nos valores das estimativas e aumento no tamanho dos ICs nos bancos BI-5 e BI-20, devido à redução do tamanho amostral. Portanto, restringir a análise ao conjunto de casos que têm observações completas pode levar a conclusões erradas [2,6,13,17].

Não houve muita diferença entre os modelos obtidos com os diferentes métodos de IM. Isso pode ser justificado pelo fato de que somente uma variável teve seus dados imputados [6].

Freqüentemente, em estudos epidemiológicos, modelos são estimados pela análise de regressão logística, e uma abordagem comum é restringir-se à análise dos casos completos. Essa abordagem exclui todos os pacientes que tenham a informação incompleta em qualquer um dos preditores. Tais modelos podem conter

Tabela 5

Estimativas da regressão logística após imputações múltiplas pelo método BLR em diferentes regressões. Mecanismo MCAR (BI-20, n = 383).

Variáveis independentes	RC [IC95%] e (erro padrão) dos modelos logísticos ajustados			
	Banco completo	IM(1) *	IM(2) **	IM(3) ***
ASA III	3,4 [1,5; 7,9]	3,4 [1,5; 7,9]	3,5 [1,5; 8,1]	3,4 [1,5; 7,9]
	-0,422	-0,422	-0,423	-0,424
ASA IV/V	20,2 [8,8; 46,0]	20,2 [8,8; 46,1]	20,3 [8,9; 46,6]	19,6 [8,5; 45,2]
	-0,421	-0,423	-0,423	-0,425
Idade ≥ 75	2,9 [1,5; 5,8]	2,9 [1,5; 5,8]	2,9 [1,5; 5,8]	2,9 [1,5; 5,8]
	-0,348	-0,35	-0,351	-0,351
Albumina até 2,2g/dl	5,3 [2,7; 10,5]	5,2 [2,5; 10,4]	5,1 [2,5; 10,3]	5,5 [2,7; 11,2]
	-0,349	-0,36	-0,362	-0,363
Albumina 2,3 a 3,0g/dl	2,1 [1,1; 4,1]	2,0 [1,0; 4,1]	1,9 [0,9; 3,9]	2,0 [1,0; 4,1]
	-0,345	-0,359	-0,361	-0,363

* IM(1): Albumina = β_1(ASA III)+β_2(ASA IV/V)+β_3(Idade ≥75)+cte;

** IM(2): Albumina = β_1(ASA III)+β_2(ASA IV/V)+β_3(Cirurgia urgente)+cte;

*** IM(3): Albumina = β_1(ASA III)+β_2(ASA IV/V)+β_3(Idade ≥75)+β^4(Cirurgia urgente)+cte.

coeficientes menos fidedignos e as estimativas podem ser viesadas se grupos homogêneos de pacientes forem excluídos da análise. Em conseqüência disso, tem sido recomendado que os dados faltantes sejam imputados por métodos apropriados antes de se fazer as análises [13,17].

Em todos os modelos ajustados com dados imputados, a categoria "2,3 a 3,0 g/dl" da variável albumina não foi significativa, enquanto que, na análise com os dados completos, esta categoria mostrou-se significativa. Uma justificativa poderia ser que, quando observados os resultados da estimação pontual e da estimação por intervalo para a RC, vê-se que a associação desta variável com o desfecho foi bastante fraca, pois RC = 2,1 e IC95%: 1,1; 4,1 e, além disto, como a variância da IM leva em conta a variância dentro das imputações e também a variância entre imputações, é natural que a variância estimada seja um pouco maior do que a com dados completos. Portanto, como a estimativa pontual do coeficiente obtida pela IM é bem próxima do valor real, mas a variância é um pouco maior, isto faz com que o IC seja um pouco mais largo.

Além do mais, analisar somente os casos completos, isto é, sem imputação, pode resultar em tamanhos de amostras menores que o planejado e ainda gerar modelos com menos preditores do que o caso do banco original (verdadeiro). Portanto, uma justificativa para o uso da imputação de dados é que quando se tem perda de dados o poder estatístico diminui, pois diminui o tamanho da amostra. Embora já existam na literatura alternativas para ajustar modelos com dados faltantes sem fazer imputação, tais como métodos não interativos, maximização direta da verossimilhança dos dados observados ou métodos *bootstrap*, há o problema de que não se encontram facilmente estas alternativas implementadas computacionalmente [18].

Para a análise de regressão logística como feita neste trabalho, quando não há informação em alguma das variáveis do modelo, o sujeito é inteiramente retirado da análise. Portanto, a amostra torna-se menor do que a planejada inicialmente. A amostra completa aqui utilizada era de 470 pacientes. Essa amostra tinha um poder de 80% para detectar um RC de 2,78 e 90% para detectar um RC de 3,14. Quando foram excluídos 5% da amostra, ficando-se com n = 440, esse poder caiu para 77% e 87,8%, respectivamente, e quando se excluiu 20% da amostra (n = 383), o poder ficou em 70,2% e 82,3%, respectivamente. Esses valores reforçam a importância da imputação de dados.

Neste trabalho, o mecanismo da não-resposta foi MCAR pela maneira como a perda foi gerada por simulação, ou seja, a probabilidade de que os dados da albumina fossem faltantes não dependia de nenhuma das variáveis observadas. Isso pode ter afetado os resultados, pois alguns autores recomendam que quando os dados faltantes são MAR, e as variáveis das quais a probabilidade de perda do dado faltante são bem identificadas, melhora o desempenho da IM, pois estas variáveis podem ser incluídas no modelo de imputação. A suposição MAR é a mais usada

nos estudos epidemiológicos, não porque seja mais plausível na prática, mas porque representa a condição mais geral sob a qual inferências válidas podem ser obtidas sem se fazer referência ao mecanismo de não-resposta [8,19]. Entretanto, pode ser citado o trabalho de Moons et al. [20], que simularam dados faltantes em um conjunto de dados reais, gerando amostras sob as suposições MCAR e MAR, obtendo os mesmos resultados com as duas amostras.

Observando-se os resultados deste trabalho sob o prisma de inclusão de variáveis na regressão linear para obter valores preditos para a imputação, verifica-se que a inclusão de um número maior de variáveis no momento da imputação não necessariamente melhora o ajuste do modelo feito com o banco completo. Os resultados obtidos por todas as IM foram bastante parecidos, mesmo quando se incluiu mais variáveis, caso da IM(3).

É interessante que se tenha um guia para a seleção das variáveis que entrarão para a imputação. Se muitas variáveis com potencial para imputação estiverem disponíveis, deve-se estabelecer um procedimento formal para a seleção das variáveis, algo como o "guia" apresentado neste artigo [16]. Por causa da natureza Bayesiana da IM, no caso de super-ajuste, isto é, incluir preditores redundantes, pode-se esperar redução de precisão nas estimativas finais, mas não outros problemas, como viés. Em contraste a isso, a omissão de importantes preditores da perda pode gerar viés. Assim, pode ser melhor super-ajustado do que subajustado [8].

A partir deste único trabalho não é possível tirar-se conclusões sobre qual dos métodos de IM usados é mais apropriado para se lidar com dados faltantes, pois os resultados foram bastante semelhantes. Entretanto, é possível afirmar que é melhor imputar do que analisar somente os casos completos. O que diferencia um método do outro é que no método PMM há um componente "hot deck" em sua aplicação, ou seja, no método PMM todos os valores imputados são valores observados na amostra, enquanto isto não acontece no método BLR [2]. Mas ficou claro que isso praticamente não influenciou os resultados, dando a idéia de que qualquer um dos dois métodos pode ser usado. Ainda, pode-se afirmar que, teoricamente, a IM tem vantagens sobre a imputação única [6].

Para este trabalho foi utilizado o pacote MICE no ambiente do aplicativo R para se fazer a IM. Porém, nos últimos anos, vários aplicativos, livres e comerciais, implementaram em suas rotinas técnicas de imputação, tanto imputação única como IM. A literatura recente traz discussões acerca dessas implementações [10,11,12].

Historicamente e por razões práticas, para IM se usava m pequeno, como valores entre 3 e 10. Usualmente $m = 5$ é o mais freqüente, sendo que este foi o valor utilizado neste trabalho por ser o *default* do MICE. Por resultados teóricos de Rubin sabe-se que esses valores sugeridos para m são suficientes para que as conclusões sejam válidas. No entanto, hoje em dia, com os avanços computacionais, tornou-se praticável que o número de m de imputações seja muito maior sem que isto cause problemas. É possível que se use m igual a 100 ou 200 [8,21].

Como conclusão, é recomendado que os pesquisadores ao analisarem seus dados não ignorem simplesmente o problema de dados faltantes. Imputar dados faltantes pode aumentar consideravelmente a confiabilidade dos resultados obtidos. Além disso, estratégias para se lidar com dados faltantes podem aumentar o tamanho efetivo do conjunto de dados, tornando as análises mais poderosas [13].

Um aspecto interessante da IM é a combinação do paradigma Bayesiano no passo da imputação e a abordagem freqüentista no final da análise dos dados [8].

Finalmente, sugere-se que outros estudos empíricos com mais variáveis com dados faltantes e maiores proporções de dados faltantes devem ser feitos para mostrar o comportamento dos resultados dos diferentes métodos de imputação múltipla. Para ajudar os pesquisadores da área médica, mais trabalhos com foco na metodologia devem ser produzidos, indicando que deve-se usar técnicas de imputação para tratar o problema de dados faltantes, e ressaltando as vantagens da IM sobre a imputação única [6].

Resumo

Em situações com dados faltantes, é comum restringir-se à análise dos sujeitos com dados completos. Porém, as estimativas com apenas esses sujeitos podem tornar-se viesadas. A prática de preenchimento de dados faltantes é a chamada técnica de imputação. Este trabalho tem como objetivo divulgar o método de imputação múltipla. Em um conjunto de dados de 470 pacientes cirúrgicos, foram ajustados modelos logísticos para o desfecho óbito. Foram gerados dois conjuntos de dados incompletos: um com 5% e outro com 20% de dados faltantes para uma variável. Foram ajustados modelos para o conjunto completo, com dados faltantes e para o conjunto completado por imputação múltipla. As estimativas obtidas pela análise dos conjuntos com dados faltantes e com o conjunto completo foram diferentes, principalmente as do conjunto com 20% de dados faltantes. A imputação múltipla utilizada pareceu eficiente, pois os resultados conseguidos com o banco completado por imputações foram próximos dos obtidos com o conjunto completo. Porém, um coeficiente deixou de ser estatisticamente significativo. A imputação múltipla se mostrou superior à análise do conjunto com dados faltantes, que desconsiderou os casos incompletos.

Interpretação Estatística de Dados; Modelos Estatísticos; Base de Dados

Colaboradores

L. N. Nunes participou da concepção do estudo, da análise dos dados, da discussão dos resultados, da redação e da revisão do manuscrito. J. M. G. Fachel participou da concepção do estudo, da redação e da revisão do manuscrito. M. M. Klück contribuiu na concepção do estudo e revisão do manuscrito.

Referências

1. Rubin DB. Multiple imputation after 18+ years. J Am Stat Assoc 1996; 91:473-89.
2. Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley; 1987.
3. Little RJA. Regression with missing Xs – a review. J Am Stat Assoc 1992; 87:227-37.
4. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res 1999; 8:3-15.
5. Zhang P. Multiple imputation: theory and method. Int Stat Rev 2003; 71:581-92.
6. 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; 59:1102-9.
7. White IA, Wood A, Royston P. Editorial: multiple imputation in practice. Stat Methods Med Res 2007; 16:195-7.
8. Kenward MG, Carpenter J. Multiple imputation: current perspectives. Stat Methods Med Res 2007; 16:199-218.
9. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002; 7:147-77.
10. Horton NJ, Lipsitz SR. Multiple imputation in practice: comparison of software packages for regression models with missing variables. Am Stat 2001; 55:244-54.
11. Acock AC. Working with missing values. J Marriage Fam 2005; 67:1012-28.

12. Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. Am Stat 2007; 61:79-90.
13. Harrell Jr. FE. Regression modeling strategies: with applications to linear models, logistic regression and survival analysis. New York: Springer-Verlag; 2001.
14. Klück M. Metodologia para ajuste de indicadores de desfechos hospitalares por risco prévio do paciente [Tese de Doutorado]. Porto Alegre: Faculdade de Medicina, Universidade Federal do Rio Grande do Sul; 2004.
15. van Buuren S, Oudshoorn CGM. Multivariate imputation by chained equations. MICE V1.0 user's manual. Leiden: TNO Preventie en Gezondheid; 2000.
16. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med 1999; 18:681-94.
17. 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; 16:277-98.
18. Meng X-L. Missing data: dial M for ???. J Am Stat Assoc 2000; 95:1325-30.
19. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clinical Epidemiol 2006; 59:1087-91.
20. Moons KG, Donders RA, Stijnen T, Harrell Jr. FE. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 2006; 59:1092-101.
21. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prev Sci 2007; 8:206-13.

Recebido em 14/Fev/2008
Versão final reapresentada em 31/Jul/2008
Aprovado em 04/Set/2008