UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

TESE DE DOUTORADO
MÉTODOS DE IMPUTAÇÃO DE DADOS APLICADOS NA ÁREA DA SAÚDE

Luciana Neves Nunes

Orientador: Profa. Dra. Jandyra Maria Guimarães Fachel

Porto Alegre, setembro de 2007

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

TESE DE DOUTORADO
MÉTODOS DE IMPUTAÇÃO DE DADOS APLICADOS NA ÁREA DA SAÚDE

Luciana Neves Nunes

Orientadora: Profa. Dra. Jandyra Maria Guimarães Fachel

A apresentação desta tese é exigência do Programa de Pós-graduação em Epidemiologia, Universidade Federal do Rio Grande do Sul, para obtenção do título de Doutor.

Porto Alegre, Brasil.
2007

BANCA EXAMINADORA

Prof. Dr. João Riboldi, Programa de Pós-graduação em Epidemiologia, Universidade Federal do Rio Grande do Sul.

Prof. Dra. Maria Teresa Anselmo Olinto, Programa de Pós Graduação em Saúde Coletiva, Universidade do Vale do Rio dos Sinos.

Prof. Dra. Sidia Maria Callegari-Jacques, Departamento de Estatística, Universidade Federal do Rio Grande do Sul.

Missing data are not merely a nuisance during data analysis. Incomplete data problems can and should inspire good scientific thinking.

Stef van Buuren and Rob Eisinga
2003

AGRADECIMENTOS

Confesso que nesse momento gostaria de ser poeta e poder escrever aqui uma poesia que pudesse transmitir o quão grata sou a tantas pessoas que fazem parte da minha vida e que me ajudaram a atingir meu objetivo de um dia ser "Doutora". Queria ser poeta para conseguir transcrever essa gratidão através da beleza das palavras. Mas como sou Estatística e Epidemiologista o que me resta é analisar meus dados, resumi-los e apresentá-los, ou seja, veja através do gráfico abaixo como eu agradeço a todos do fundo do meu coração:

Prefiro não mencionar nomes porque posso me esquecer de alguém... Todos que colaboraram sabem quem são e como fizeram isso!!!!!

SUMÁRIO

ABREVIATURAS E SIGLAS

ASA	American Society of Anestesiology
BI	Banco Incompleto
BLR	Bayesian Linear Regression
EM	Expectation Maximization
EP	Erro Padrão
HCPA	Hospital de Clínicas de Porto Alegre
IC	Intervalo de Confiança
IM	Imputação Múltipla
LOCF	Last Observation Carried Forward
MAR	Missing at Random
MCAR	Missing Completely at Random
MCMC	Markov Chain Monte Carlo
MICE	Multivariate Imputation by Chained Equations
NHANES	National Health and Nutrition Examination Surveys
NMAR	Missing Not at Random
NOCB	Next Observation Carried Backward
PMM	Preditive Mean Matching
RC	Razão de Chances
ROC	Receive Operator Caracteristic Curve
SDS	Self-reported Depression Scale
SPSS	Statistical Package for Social Science

RESUMO

Em pesquisas da área da saúde é muito comum que o pesquisador defronte-se com o problema de dados faltantes. Nessa situação, é freqüente que a decisão do pesquisador seja desconsiderar os sujeitos que tenham não-resposta em alguma ou algumas das variáveis, pois muitas das técnicas estatísticas foram desenvolvidas para analisar dados completos. Entretanto, essa exclusão de sujeitos pode gerar inferências que não são válidas, principalmente se os indivíduos que permanecem na análise são diferentes daqueles que foram excluídos. Nas duas últimas décadas, métodos de imputação de dados foram desenvolvidos com a intenção de se encontrar solução para esse problema. Esses métodos usam como base a idéia de preencher os dados faltantes com valores plausíveis. O método mais complexo de imputação é a chamada imputação múltipla. Essa tese tem por objetivo divulgar o método de imputação múltipla e através de dois artigos procura atingir esse objetivo. O primeiro artigo descreve duas técnicas de imputação múltipla e as aplica a um conjunto de dados reais. O segundo artigo faz a comparação do método de imputação múltipla com duas técnicas de imputação única através de uma aplicação a um modelo de risco para mortalidade cirúrgica. Para as aplicações foram usados dados secundários já utilizados por Klück (2004).

Palavras-chave: Métodos de imputação, imputação múltipla, dados faltantes, não-resposta.

ABSTRACT

Missing data in health research is a very common problem. The most direct way of dealing with missing data is to exclude observations with missing data, probably because the traditional statistical methods have been developed for complete data sets. However, this decision may give biased results, mainly if the subjects considered in the analysis are different of those who have been excluded. In the last two decades, imputation methods were developed to solve this problem. The idea of the imputation is to fill in the missing data with reasonable values. The multiple imputation is the most complex method. The objective of this dissertation is to divulge the multiple imputation method through two papers. The first one describes two different types of multiple imputation and it shows an application to real data. The second paper shows a comparison among the multiple imputation and two single imputations applied to a risk model for surgical mortality. The used data sets were secondary data used by Klück (2004).

Key-words: Imputation methods, multiple imputation, missing data, nonresponse.

LISTA DE QUADROS E TABELAS

LISTA DE FIGURAS

1. APRESENTAÇÃO

Este trabalho consiste na tese de doutorado intitulada "MÉTODOS DE IMPUTAÇÃO DE DADOS APLICADOS NA ÁREA DA SAÚDE", apresentada ao Programa de Pós-Graduação em Epidemiologia da Universidade Federal do Rio Grande do Sul, em 14 de setembro de 2007. O trabalho é apresentado em três partes, na ordem que segue:

1. Introdução, Objetivos e Revisão da Literatura.
2. Artigos
3. Conclusões e Considerações Finais.

No anexo consta o Projeto de Pesquisa.

2. INTRODUÇÃO

Uma complicação comum em investigações científicas é a ocorrência de dados faltantes ou dados perdidos (*missing data*), especialmente na área da Saúde e das Ciências Sociais (Rubin, 1996). Determinar a abordagem analítica adequada para bancos 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 conjunto de dados. O desenvolvimento de métodos estatísticos direcionados a solucionar problemas de dados faltantes tem sido uma área de pesquisa bastante ativa nas últimas décadas (Rubin, 1987; Little, 1992; Schafer, 1999; Zhang, 2003; Van der Van der Heijden, 2006; Harel, 2007; Kenward, 2007).

Em situações com dados faltantes, uma abordagem bastante comum é restringir a análise aos sujeitos com dados completos nas variáveis envolvidas. Porém, as estimativas obtidas com tais análises podem ser viesadas, especialmente se os indivíduos que são incluídos na análise são sistematicamente diferentes daqueles que foram excluídos em termos de uma ou mais variáveis. Para contornar esse problema, desde os anos 80 surgiram técnicas estatísticas que envolvem a substituição dos dados faltantes por estimativas de valores plausíveis a serem "imputados" aos dados faltantes. Esta técnica denomina-se imputação de dados faltantes na literatura estatística e seu uso vem generalizando-se e estendendo-se a outras áreas (Rubin, 1996; Schafer, 2002, Fraser, 2007).

Essas técnicas têm por objetivo "completar" os bancos de dados 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 ou pela mediana da variável, por interpolação ou até por

regressão linear. Todas essas técnicas mencionadas permitem "preencher" os dados faltantes através 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. Com a idéia de se resolver essa questão foi desenvolvida (Rubin,1987) a técnica de Imputação Múltipla (IM).

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 (Harel, 2007).

A literatura sobre IM tem tido uma expansão considerável desde o início da década de 90. Entretanto, esse número é grande no que diz respeito a publicações com aplicações. Mais textos metodológicos têm que ser desenvolvidos e divulgados antes que os pesquisadores possam usar a IM rotineiramente e com confiança (White, 2007).

É possível se detectar o crescente interesse por essa área através de uma breve busca na internet. No PubMed, uma busca com a palavra chave "imputation" indicou 571 trabalhos publicados (01/09/2007), sendo que só nos últimos 12 meses foram 66 trabalhos. São 300 trabalhos com a palavra chave "multiple imputation", 192 publicações com as palavras "multiple imputation" e "missing data", 46 com as palavras "multiple imputation" e "logistic regression" e 13 com as palavras "multiple imputation" e "cox".

Ainda, importantes periódicos perceberam a necessidade de fazer edições especialmente dedicadas ao assunto de imputação. Por exemplo, a revista *Statistics in*

Medicine, dedicou os fascículos 1 a 3 do volume 16 (1997) para divulgar métodos para tratamento de dados faltantes. Em 2003, a revista *Statistica Neerlandica* também dedicou um fascículo inteiro para a imputação de dados por considerar um assunto de extrema relevância na atualidade. O periódico *Journal of Clinical Epidemiology* publicou uma série especial de artigos originais tendo como assunto o tratamento de dados faltantes, no volume 59 de outubro de 2006. Recentemente, a revista *Statistical Methods in Medical Research* de junho de 2007 dedicou sua edição a artigos sobre IM que surgiram a partir de um workshop em IM realizado em *Cambridge, UK (MRC Biostatistics Unit)* nos dias 12 e 13 de abril de 2005. Estes fatos revelam que o estudo de metodologias para dados faltantes vem sendo bastante debatidos, o que indica a pertinência desse 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.) (White, 2007).

Desde sua introdução há mais ou menos 30 anos atrás na análise de pesquisas, a IM se tornou uma abordagem importante e influente na análise de dados incompletos. Durante esse período, a amplitude 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 (Kenward, 2007).

A perda de dados é um grande desafio no planejamento e análise dos estudos epidemiológicos. Por exemplo, nestes estudos freqüentemente o objetivo é determinar

variáveis preditoras que contribuem para predizer a ausência ou presença de uma doença em uma população. Perda de informações, tanto nos preditores como no desfecho, pode levar a problemas sérios para a análise dos dados. Portanto, é importante que se estabeleçam estratégias para lidar com dados faltantes nas variáveis, seja planejando a pesquisa com o máximo de esforço, para evitar a perda de informações, seja abordando os dados faltantes com técnicas adequadas desenvolvidas para contornar esse problema (van der Heijden, 2006).

A proposta desse trabalho é a divulgação dos métodos de imputação para ajudar os pesquisadores da área da saúde em suas futuras análises, focando principalmente a técnica da Imputação Múltipla. Através da revisão de literatura e apresentação de dois artigos pretende-se defender a idéia de que, quando o pesquisador defrontar-se com a situação de dados faltantes, considere seriamente a possibilidade de imputar os dados, pois obterá melhores resultados do que simplesmente ignorar os dados faltantes. Segundo Donders (2006), imputar dados é simples e deve ser utilizado para tratar dados faltantes.

Mais especificamente, os objetivos desse trabalho estão descritos no próximo capítulo.

3. OBJETIVOS

O objetivo geral desse trabalho é divulgar a metodologia de imputação de dados, principalmente a imputação múltipla, para pesquisadores da área da saúde. Essa divulgação será feita através de breve descrição de métodos de imputação única e um detalhamento maior da imputação múltipla .

São objetivos específicos desse trabalho:

Artigo 1:

- Descrever a imputação múltipla .
- Comparar duas técnicas de imputação múltipla a partir de uma aplicação a dados reais.

Artigo 2:

- Comparar os resultados obtidos com diferentes técnicas de imputação: imputações únicas e imputação múltipla a partir da análise de um conjunto de dados reais.

Objetivos específicos comuns aos dois artigos:

- Mostrar de forma didática como podem ser implementados métodos de imputação à análise de dados faltantes.
- Mostrar como são utilizadas rotinas de imputação múltipla através do pacote MICE no R.

4. REVISÃO DE LITERATURA

Com ou sem dados faltantes, quando se analisa um conjunto de dados, o objetivo de um procedimento estatístico deve ser fazer inferências válidas e eficientes para uma população de interesse. Tentativas de recuperar dados perdidos sem um estudo adequado podem prejudicar a inferência. Por exemplo, a prática comum de se usar a substituição pela média, ou seja, substituir os dados faltantes pela média geral dos valores observados, pode predizer o dado faltante de forma aceitável, mas distorcer as estimativas da variância, do erro padrão e da correlação (Schafer e Graham, 2002).

Critérios básicos para avaliar procedimentos de estimação foram estabelecidos por Neyman e Pearson (1933) e Neyman (1937). Seja Q uma quantidade populacional a ser estimada e seja $\hat{Q}$ uma estimativa de Q. Se a amostra contém dados faltantes, então o método para lidar com eles deve levar em conta o procedimento estatístico que será usado para calcular $\hat{Q}$. Se o procedimento funciona bem, então $\hat{Q}$ estará perto de Q tanto na média, para amostras repetidas, como para uma amostra em particular. Isto é, deseja-se que o viés – a diferença entre a média de $\hat{Q}$ e o verdadeiro valor Q – seja pequeno e também se deseja que a variância ou desvio padrão de $\hat{Q}$ seja pequeno. (Schafer e Graham, 2002).

Quando dados faltantes ocorrem por razões fora de controle, devem-se fazer suposições acerca do processo que os gerou. Sugere-se que as suposições devem ser explícitadas e a sensibilidade à mudança nos resultados por violação dessas suposições deve ser investigada (Schafer e Graham, 2002).

Uma característica da análise de dados incompletos é que ela depende de suposições que não podem ser testadas. Nesse sentido, tais análises pertencem a uma classe na qual se incluem: confundidores não observados, erros de medida e não adesão a tratamento. Como conseqüência, existe uma concordância geral de que a análise deve ser repetida sob variadas suposições, para ver se as conclusões são sensíveis a essas mudanças (Kenward, 2007).

4.1 "Não-resposta" ou "dados faltantes" (*missing data*)

Em muitas pesquisas, algumas das unidades que são contatadas podem não responder a algum item ou a alguns itens da pesquisa. Tal não-resposta que gera o chamado dado faltante é comum acontecer na prática, não importando se as unidades de pesquisa são indivíduos, domicílios, empresas, escolas, etc. O problema criado pela não-resposta é que, freqüentemente, os objetivos pretendidos pela pesquisa não são alcançados, gerando menos eficiência nas estimativas por causa da redução do tamanho da amostra e, em alguns casos, métodos tradicionais de análise de dados completos não podem ser usados diretamente, devido a magnitude do número de dados faltantes. Além disso, possíveis vieses podem existir por causa da diferença que freqüentemente existe entre respondentes e não-respondentes. Particularmente, é muito difícil eliminar esses vieses porque geralmente não se conhecem as razões precisas para a existência da não-resposta (Rubin, 1987).

Por exemplo, pesquisas realizadas com dados secundários coletados a partir de prontuários de pacientes de hospitais podem ter dados faltantes no banco de dados. Essa falta de dados pode ser decorrente de vários motivos. Especificamente, pode-se citar uma pesquisa realizada no HCPA (Klück, 2004) em que alguns prontuários não continham a informação do resultado laboratorial da Albumina. Os motivos levantados

foram principalmente dois: o médico que preenchia o prontuário não via importância em colocar tal informação e não havia obrigatoriedade de preenchimento desses dados. A partir do trabalho realizado por Klück (2004), o HCPA passou a ter como campo de preenchimento obrigatório nos prontuários de pacientes que serão submetidos à cirurgia o valor da Albumina obtido em exames laboratoriais pré-operatórios.

4.2 Mecanismos de não-resposta

Como a ausência de dados é praticamente inevitável, na etapa de delineamento do estudo deve-se planejar a investigação de potenciais preditores dos valores faltantes e utilizá-los nas análises. É importante que os dados faltantes não sejam considerados apenas como um problema de análise de dados, mas também como uma questão de planejamento da pesquisa e da interpretação dos resultados. É também importante que os mecanismos que geram os dados faltantes sejam identificados e levados em conta nas análises dos dados. Partindo da idéia de que a presença de dados faltantes é um problema que freqüentemente acontece na prática, Rubin (1976) faz uma revisão sobre os mecanismos de não-resposta e os métodos de inferência que se utilizam na presença de tal modelo.

A principal conclusão a que o autor chega é que, na prática, o pesquisador deve considerar o mecanismo que gera a não-resposta de forma mais freqüente do que usualmente é feito. Entretanto, para fazer isso, o pesquisador precisa de modelos para esses processos de não-resposta que não têm tido atenção na literatura médica. Os tipos e mecanismos de não-resposta citados por Rubin e utilizados ao longo da literatura sobre dados faltantes são:

a) ***Dado faltante completamente aletório - MCAR (Missing Completely at Random)***

b) ***Dado faltante aleatório - MAR (Missing at Random)***

c) ***Dado faltante não aleatório - NMAR (Missing Not at Random)***

4.2.1 – Faltante completamente aleatório – MCAR *(Missing Completely at Random)*

Dados são faltantes completamente ao acaso quando as razões para as perdas não são relacionadas a quaisquer respostas dos sujeitos, incluindo o valor faltante. Esse é o mecanismo mais restritivo, pois impõe que a probabilidade de não-resposta seja a mesma para diversas situações. Assim, em um delineamento tipo "follow-up", somente uma amostra aleatória de 50% dos indivíduos arrolados inicialmente são seguidos, não há problema, pois esse grupo será uma representação não viesada da população.

4.2.2 – Faltante aleatório – MAR *(Missing at Random)*

Dados faltantes são considerados MAR quando o padrão de perda em uma variável é previsível a partir de outras variáveis no banco de dados e não é devido à variável específica na qual os dados são perdidos. Por exemplo, considere uma pesquisa na qual as mulheres são menos propensas a fornecer sua renda pessoal. Se conhecermos o sexo de todos os sujeitos e tivermos a renda para algumas mulheres, estimativas não viesadas da renda podem ser feitas. Isto porque a renda que se tem de algumas mulheres é uma amostra aleatória das rendas de todas as mulheres.

4.2.3 – Faltante não aleatório – NMAR *(Missing Not at Random)*

O dado faltante é não aleatório se está relacionado com os valores não observados, mesmo se controlado para outras variáveis na análise. Dados que são mais propensos a serem faltantes, em geral, são aqueles situados nos extremos da

distribuição, com valores mais altos ou mais baixos do que o padrão da amostra. Um exemplo é quando sujeitos com níveis de renda muito baixos ou muito altos têm probabilidade menor de responder sobre sua renda pessoal numa entrevista.

Dados MCAR e MAR também são chamados de dados faltantes ignoráveis, enquanto os NMAR são os não-ignoráveis. Esse último tipo é o mais difícil de se tratar numa análise, enquanto o MCAR é o mais fácil. A habilidade em se analisar corretamente os dados MAR depende da disponibilidade de variáveis auxiliares. Muitos dos métodos disponíveis para tratar dados faltantes assumem que os dados são MAR (Harrell, 2001).

Em geral, os dados faltantes não são nem MCAR nem NMAR. Usualmente, a probabilidade de que um dado seja faltante depende de outras variáveis observadas em relação ao sujeito, isto é, a razão para a perda da informação pode ser baseada em outras informações observadas. Esse mecanismo de não-resposta é confusamente chamado de "faltante aleatório" (MAR) porque o dado faltante pode de fato ser considerado ao acaso, mas condicional às outras variáveis observadas que determinam essa perda e estão disponíveis para a análise (Rubin, 1976; Donders, 2006). Por exemplo, suponha que queremos avaliar o valor preditivo de um teste diagnóstico particular e que os resultados são conhecidos para todos os sujeitos doentes, mas desconhecido para uma amostra aleatória de sujeitos que não têm a doença. No caso, o mecanismo de não-resposta seria MAR, pois, condicionando o resultado do teste em relação à variável observada (aqui, ter ou não a doença), a perda é aleatória. É claro que não é necessário que o dado faltante dependa do desfecho, aqui está um exemplo da situação mais simples em que se tem um preditor e um desfecho (Donders, 2006).

A suposição MAR ocupa uma importante posição na estrutura de dados faltantes, 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 (Kenward, 2007).

Quando os dados são NMAR é necessário que se incorpore explicitamente o mecanismo de não-resposta, algo que na maioria das situações é desconhecido. Embora a suposição MAR torna a análise dos dados faltantes mais simples, infelizmente, esta é uma suposição que não pode ser testada como se faz com a suposição de normalidade (Kenward, 2007). Desta forma, é muitas vezes impossível determinar se os dados faltantes são MAR ou NMAR, podendo se apenas hipotetizar (van der Heijden, 2006).

Embora a suposição MAR não possa ser testada, tem sido apontado na literatura que se incluirmos variáveis suficientes no modelo de imputação, mais próximo se chega dessa suposição (Schafer, 1997; Harel, 2007). Adicionalmente, uma estimação eficiente com dados faltantes não ignoráveis (NMAR) requer um bom conhecimento prévio do mecanismo que gerou a não-resposta, pois os dados não contêm informações sobre qual modelo não ignorável seria adequado. Por isso, os resultados seriam sensíveis aos modelos assumidos (Harel, 2007).

4.3. Padrões de não-resposta

4.3.1 - Monotônicos

Muitos conjuntos de dados podem ser arranjados na forma de matriz, onde as linhas são os indivíduos e as colunas são as variáveis. Com esse tipo de arranjo é possível identificar padrões de resposta como indicada na Figura 1. Quando ocorrem dados faltantes em somente uma das variáveis (o dado faltante pode ocorrer tanto nas variáveis preditoras como na variável resposta), configura-se o padrão univariado (Figura 1a) que é um caso especial de padrão monotônico. Quando há dados faltantes em mais do que uma variável e é possível organizar os dados conforme a figura 1b, ou

seja, as colunas podem ser arranjadas de tal forma que X_{j+1} é observado para todo caso que é observado em X_j, temos o padrão monotônico (Rubin, 1987; Schafer e Graham, 2002; Little, 1992).

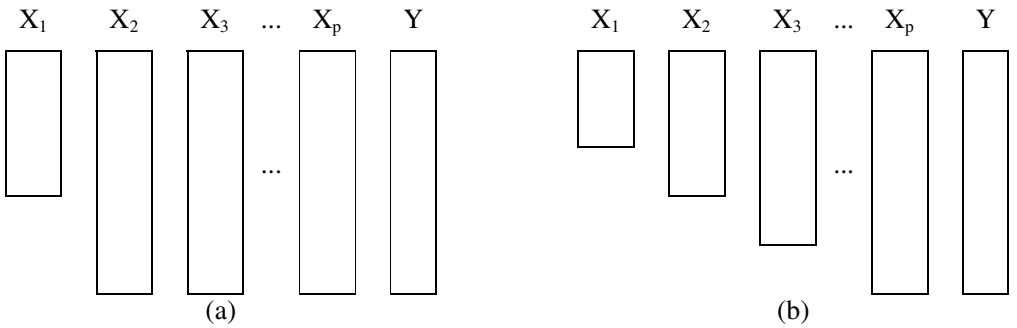

Figura 1 – Padrão monotônico de não-resposta

4.3.2 – Não-Monotônicos

A Figura 2a mostra um padrão onde duas variáveis (por exemplo, X_1 e X_2) nunca são observadas juntas. Tais dados originam-se quando duas amostras contendo observações em X_1 e Y e X_2 e Y são unidas em um só banco de dados.

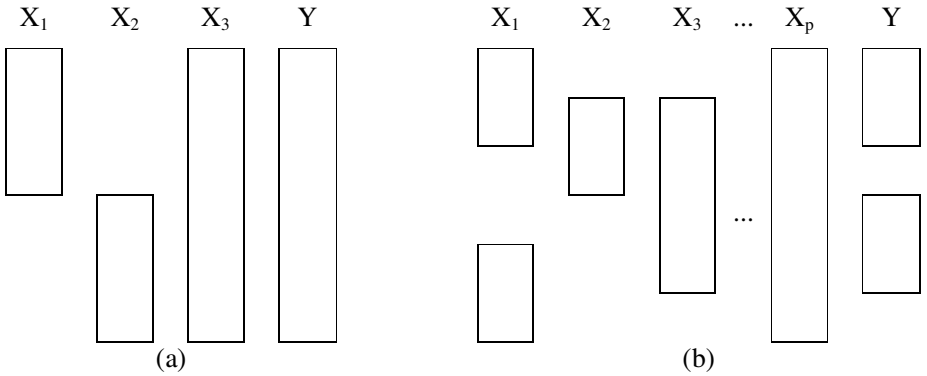

Figura 2 – Padrão não-monotônico de não-resposta

As estimativas a partir desse padrão requerem suposição (explícita ou implícita) sobre a associação condicional de X_1 e X_2 dado X_3 e Y. E por fim, a figura 2b

representa um padrão geral sem estrutura também conhecido como padrão arbitrário (Rubin, 1987; Schafer e Graham, 2002; Little, 1992).

4.4 Imputação

Nas duas últimas décadas, vários métodos de inferência estatística quando os dados são incompletos têm sido desenvolvidos, tais como o método EM (Expectation-Maximization) (Dempster, Laird e Rubin, 1977; Little e Rubin, 1987) ou os modelos de efeitos aleatórios (McCullagh e Nelder, 1989; Diggle et al., 1994). Esses métodos são baseados na estimativa de máxima verossimilhança e, sob certas suposições, geram inferências válidas dos parâmetros desconhecidos. Entretanto, podem surgir complicações computacionais e a precisão das inferências geralmente fica diminuída por causa da presença de dados faltantes (Zhang, 2003).

Outra possível abordagem é preencher os dados faltantes com valores plausíveis e então aplicar métodos tradicionais de análise de dados completos para fazer inferências válidas e eficientes. Essa abordagem é chamada de imputação de dados e evita a complexidade gerada pelos dados faltantes. Entretanto, as incertezas associadas à imputação precisam ser tratadas apropriadamente porque os dados imputados não são dados reais. Três tipos de incertezas estão associados ao processo de imputação: (a) aquelas devidas à modelagem da distribuição conjunta das variáveis respostas e as indicadoras de dados faltantes; (b) aquelas devidas à amostragem (seleção aleatória de valores a serem imputados) a partir de um modelo de imputação assumindo que os dados observados e os valores dos parâmetros do modelo são conhecidos e (c) as incertezas sobre os valores dos parâmetros do modelo (Zhang, 2003).

Segundo Tang et al. (2005), a imputação tem se tornado uma estratégia comum para se lidar com dados incompletos em pesquisas tanto na área da Saúde como em outras áreas. Com a imputação múltipla aparecendo como um importante método usado

pelos estatísticos para, além de fornecer as estimativas pontuais, apresentar também a precisão das estimativas a partir do que é gerado pela incerteza dos dados, quando os dados faltantes são substituídos. Uma característica importante das técnicas de imputação é que elas permitem aos investigadores que usem diretamente estratégias de análise de bancos de dados completos.

4.5 – Métodos de imputação única

Métodos de imputação simples ou única são métodos de substituição de dados faltantes quando os dados perdidos são substituídos uma única vez por algum dos métodos citados a seguir (Engels, 2003):

4.5.1 Substituição por um valor de tendência central

Os dados faltantes das variáveis quantitativas são substituídos pela média da variável. Pode ser a média geral, ou seja, a média dos valores observados, ou a média de um grupo mais similar ao do caso com dado faltante, identificado por uma ou mais variáveis categóricas presentes no banco de dados. Também se podem substituir os dados faltantes pela mediana da variável ou pela mediana de um grupo de casos mais similares. Sempre que existirem valores extremos (*outliers*) na amostra, é recomendado utilizar o valor da mediana ao invés do valor da média. Se a variável com dados faltantes é categórica ordinal, utiliza-se a mediana ou pode-se também utilizar o valor modal para substituição do dado faltante. Se a variável é categórica não ordinal, é recomendado utilizar a moda e se não houver moda, sorteia-se uma categoria das com maior freqüência.

4.5.2 “Hot deck”

Os valores de respondentes, que são similares em relação a variáveis auxiliares, são selecionados para a imputação. São os chamados “doadores”. Ou seja, localiza-se o indivíduo com dado observado mais parecido com o indivíduo com dado faltante em relação às variáveis auxiliares e substitui-se o dado faltante pelo valor do respondente pareado. Se houver mais de um respondente pareado, é usado o método de imputação do “vizinho mais próximo”, onde algum critério de classificação é desenvolvido para determinar o registro mais semelhante àquele com o dado faltante e esse registro se torna doador desses dados. Por exemplo, em uma pesquisa sobre depressão em que o critério diagnóstico é fechado pelo preenchimento de um instrumento validado, em que os itens têm cinco categorias de respostas, um item não preenchido pode ser imputado da seguinte maneira: cria-se uma matriz de “doadores” a partir dos itens preenchidos e de outras variáveis auxiliares, tipo sexo e faixa etária, ou seja, uma matriz de padrões de respostas. Verifica-se qual indivíduo respondente tem o mesmo padrão do não-respondente em relação a sexo e faixa etária. Aquele de mesmo padrão é o doador, isto é, o dado faltante no item será preenchido com a resposta do doador.

4.5.3 Regressão (média predita)

Os valores imputados são preditos através de regressão simples ou múltipla, que pode ser usada simplesmente utilizando uma ou mais variáveis existentes para predizer os valores faltantes de outra variável altamente correlacionada com as anteriores. Dois tipos de regressão para imputação podem ser utilizados: regressão e regressão com termo de adição da variância do erro. A imputação que usa a regressão faz com que indivíduos que têm os mesmos valores nas mesmas covariáveis fiquem com o mesmo valor imputado, pois o valor predito é o mesmo. Para resolver esse problema, pode ser

usada a regressão que considera o erro aleatório em que é adicionado ao valor predito um valor escolhido ao acaso de uma distribuição $N(0, s_e^2)$, onde s_e^2 é a variância residual da regressão.

4.5.4 Estimativa de Máxima Verossimilhança

Esse método faz referência ao algoritmo EM (Expextation-Maximization) e é atualmente um método bastante comum de imputação. O algoritmo EM é utilizado quando se deseja estimar parâmetros a partir de um conjunto de dados incompletos. É um processo iterativo em que se repetem dois passos até que haja convergência: E (Estimação) e M (Maximização). No passo E se estima os dados faltantes para completar a matriz de dados. No passo M, com os dados completados, há uma aprendizagem das probabilidades e então essas probabilidades são usadas para se fazer a inferência no passo E, e assim, sucessivamente, o algoritmo é processado até que haja convergência.

4.5.5 Métodos de imputação única para dados longitudinais

Alguns métodos são específicos para dados longitudinais:

(a) *Média/mediana prévia na linha*: usa a média/mediana dos valores conhecidos da pessoa prévios ao dado faltante.

(b) *LOCF (Last observation carried forward) ou último valor observado*: usa o último valor observado da pessoa para substituir o dado faltante. Método comum usado para pessoas que se perdem no seguimento da diferentes etapas do estudo.

(c) *Antes/depois*: usa a média/mediana de todos os valores anteriores e posteriores ao dado faltante para cada indivíduo/observação.

(d) *NOCB (Next observation carried backward) ou próximo valor observado*: usa o próximo valor observado da pessoa para substituir o dado faltante.

(e) *Último e próximo valores*: usa a média do último e do próximo valor observados apenas.

4.6 - Imputação Múltipla (IM)

D.B. Rubin, ainda nos anos 70, propôs uma técnica chamada *imputação múltipla* (*IM*) para resolver o problema de não-resposta em pesquisas. No entanto, apenas recentemente esta técnica vem sendo mais utilizada devido aos desenvolvimentos computacionais para implementação da técnica. Essa técnica possibilita, além da estimativa pontual dos parâmetros, a inclusão da incerteza da imputação dos dados na variância dos resultados estimados, corrigindo o maior problema associado à imputação única (Schafer, 1999, Rubin, 1987).

A idéia por trás da IM é que para cada dado faltante são imputados vários valores, por exemplo, **m,** ao invés de um. Com isso, são obtidos **m** bancos de dados completos e cada conjunto de dados é analisado usando-se procedimentos para dados completos. Após, obtém-se a estimativa pontual do parâmetro que é obtida através da média das múltiplas imputações e o seu erro padrão obtido através da variância das múltiplas imputações.

A imputação múltipla divide com a imputação única duas vantagens já mencionadas: a habilidade de usar métodos de análise para bancos de dados completos e a habilidade de incorporar o conhecimento do pesquisador. Na verdade, a segunda vantagem não só se repete como é melhorada porque a IM permite que o pesquisador use seu conhecimento para refletir a incerteza sobre os valores imputados. Essa incerteza é de dois tipos: a variabilidade amostral, assumindo que as razões para a não-

resposta são conhecidas, e a variabilidade devido à incerteza sobre as razões para a não-resposta. Sob cada modelo para não-resposta, duas ou mais imputações são criadas para refletir a variabilidade amostral sob tal modelo; imputações sob mais de um modelo para não-resposta refletem a incerteza sobre as razões para a não-resposta.

Existem três vantagens importantes da IM sobre a imputação única. Primeiro, quando imputações são realizadas aleatoriamente numa tentativa de representar a distribuição dos dados, a IM aumenta a eficiência da estimação. A segunda vantagem da IM é que quando são feitas as **m** imputações sob um mesmo modelo para não-resposta, inferências válidas - isto é, que reflitam a variabilidade adicional devido aos dados faltantes sob este modelo - são obtidas simplesmente combinando inferências de dados completos de maneira simples. A terceira vantagem é que gerando imputações múltiplas sob diferentes modelos é possível um estudo da sensibilidade das inferências para vários modelos de não-resposta (Rubin, 1987).

Existem três desvantagens claras da IM sobre a imputação única. Primeiro: é necessário mais trabalho para produzir os valores a serem imputados. Segundo: mais espaço é necessário para armazenar os dados e resultados obtidos com a IM. Terceiro: é necessário mais trabalho para analisar os bancos de dados completos pela imputação múltipla do que o banco completo pela imputação única. Essas desvantagens não são sérias quando o **m** é modesto, **m** modestos são adequados quando a fração de dados faltantes é pequena. Quando a fração de dados faltantes é grande, imputações múltiplas com **m** modesto podem não ser completamente satisfatórias, entretanto a imputação única seria muito mais desastrosa (Rubin, 1987).

O esforço necessário para criar, armazenar e analisar os dados imputados por IM parece pequeno considerando-se o ganho nas inferências. Em muitos casos, esse esforço parece especialmente modesto quando comparado com o esforço necessário para

trabalhar diretamente com modelos de probabilidade explícitos para não-resposta (Rubin, 1987).

Basicamente, a *IM* consiste de três passos:

1) São obtidos **m** (em geral, **m** fica entre 3 e 10) bancos de dados completos através 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 dois são combinados de um modo simples e apropriado para se obter a chamada inferência da imputação repetida.

O primeiro passo é a parte fundamental da IM, pois as técnicas de imputação utilizadas têm que preservar a relação das observações faltantes e observadas, levar em conta o mecanismo de ausência (MCAR, MAR ou NMAR) e o padrão dos dados faltantes (monotônicos ou não-monotônicos).

4.6.1 Métodos de IM quando há padrão monotônico

Neste trabalho são detalhados dois modelos preditivos que têm como idéia geral o seguinte: quando os dados faltantes seguem um padrão monotônico de ausência, a função conjunta de verossimilhança dos dados observados pode ser fatorada num produto de funções de verossimilhança independentes, isto é,

$$L(\phi_1,\ldots,\phi_p \mid Y_{obs}) = \prod_{j=1}^{p} L(\phi_j \mid Y_{obs})$$

onde

$$L(\phi_j \mid Y_{obs}) = \prod_{j=1}^{n_j} P(y_{ij} \mid y_{i1},\ldots,y_{i,j-1},\phi_j)$$

e $P(Y_j \mid Y_i,\ldots,Y_{j-1},\phi_j)$ é a distribuição condicional de Y_j, dado $Y_1,...,Y_{j-1}$ e ϕ_j são os parâmetros. Logo, os dados faltantes com padrão monotônico podem ser imputados a

partir de distribuições independentes univariadas dado os valores prévios observados (Zhang 2003).

Outros métodos de IM quando existe padrão monotônico podem ser encontrados na literatura, entretanto fogem ao escopo deste trabalho. Ver em Rubin (1987), Schafer (1999) e Zhang (2003).

4.6.1.1 Método da Regressão Linear Bayesiana (BLR – Bayesian Linear Regresssion)

Talvez o método mais comum de predizer Y_i de um conjunto de variáveis preditoras X_i seja o modelo de regressão linear. Aqui,

$$Y_i \sim N(X_i\beta, \sigma^2)$$

é a especificação para $f(Y_i/X_i,\theta)$, $\theta=(\beta, \log\sigma)$, β um vetor de q componentes, onde q é o número de preditores, e σ um escalar. Assume-se uma distribuição *a priori* não-informativa para θ, $P(\theta)\ \alpha$ cte e para evitar-se complexidades assume-se $n_1>q$, onde n_1 é o número de respondentes.

A partir de um resultado que está em Rubin (1987) (Resultado 5.3, p.165) tem-se que a distribuição *a posteriori* de θ envolve somente as unidades com Y_i observados. Cálculos Bayesianos com o modelo de distribuição normal tais como os que aparecem em Box e Tiao (1973), por exemplo, e na página 69 do livro do Rubin (1987), mostram que *a posteriori*, σ^2 é $\hat{\sigma}_1^2(n_1 - q)$ dividido por uma variável aleatória $\chi^2_{n_1-q}$ e β dado σ^2 é normal com média $\hat{\beta}_1$ e matriz de covariância σ^2V, onde, em termos da estatística de mínimos quadrados usual baseado nos n_1 vetores (Y_i, X_i), $i \in$ observados,

$$\hat{\sigma}_1^2 = \frac{\sum_{obs}\left(Y_i - X_i\hat{\beta}_1\right)^2}{(n_1 - q)} \text{ e}$$

$$\hat{\beta}_1 = V\left[\sum_{obs} X_i^t Y_i\right], \text{ onde } V = \left[\sum_{obs} X_i^t X_i\right]^{-1}$$

Tendo a distribuição *a posteriori* de θ descrita em termos de distribuições padrão a partir das quais se pode facilmente retirar dados, a tarefa de estimação dos parâmetros a serem usados na imputação está completa.

Finalmente, a tarefa de imputação para esse modelo pode ser descrita pelos três passos a seguir:

1. Simular uma variável aleatória $\chi^2_{n_1-q}$, *g*, e seja

$$\sigma_*^2 = \frac{\hat{\sigma}_1^2 (n_1 - q)}{g}$$

2. Simular *q* variáveis independentes N(0, 1) para criar um vetor Z de *q* componentes e seja

$$\beta_* = \hat{\beta}_1 + \sigma_* [V]^{1/2} Z$$

onde $[V]^{1/2}$ é a raiz quadrada de V tal como a raiz quadrada triangular obtida pela fatoração de Cholesky.

3. Simular os n_0 valores dos $Y_{faltante}$ como

$$Y_{i*} = X_i \beta_* + z_i \sigma_*$$

onde os n_0 desvios normais z_i são simulados independentemente.

Para um novo valor a ser imputado para $Y_{faltante}$ simula-se um novo valor para o parâmetro σ_*^2. Assim, se **m** imputações são desejadas, esses três passos são repetidos **m** vezes independentemente.

4.6.1.2 Método da Média Preditiva (PMM – Predictive Mean Matching)

Este método é semelhante ao método BLR, entretanto o passo (3) é modificado da seguinte maneira:

(a) Calcula-se os n_0 valores preditos dos $Y_{faltante}$ como $Y_{i*} = X_i \beta_*$, i ∈ faltantes.

(b) Para cada Y_{i*}, i $\in$ faltantes, encontrar o respondente cujo Y_i (i $\in$ observados) seja o mais próximo de Y_{i*}. O valor do Y_i será usado para a imputação.

Esse método gera a variabilidade entre imputações desde que os passos 1 e 2 para simular β_* do método BLR e um modelo linear para guiar a escolha dos valores a serem imputados sejam utilizados. Pode-se dizer que neste método há um componente *hot-deck*, já que somente são usados para a imputação valores que foram observados.

4.6.2 Método de IM quando o padrão é não-monotônico

4.6.2.1 MCMC (Markov Chain Monte Carlo)

O método de Monte Carlo baseado em Cadeias de Markov (MCMC) tem como objetivo simular distribuições multivariadas cujo limite é uma cadeia de Markov estacionária que tem a distribuição que se deseja encontrar (Gilks, 1996). Este método não se aplica ao padrão monotônico porque a função de verossimilhança conjunta dos dados observados não pode ser fatorada em funções de verossimilhança independentes, ou seja, dados faltantes com padrão não-monotônico não podem ser imputados a partir de distribuições univariadas independentes, como é feito quando existe padrão monotônico.

Chantala e Suchindran (2005) sugerem o seguinte esquema apresentado na figura a seguir, para entender como o método MCMC funciona para a imputação:

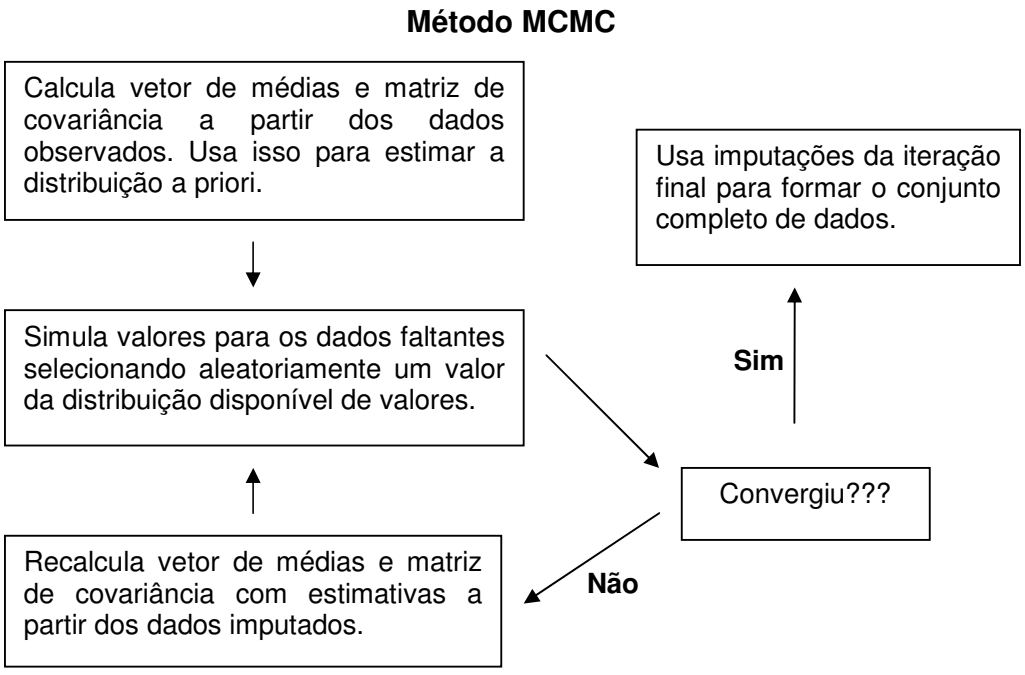

Figura 3 – Esquema do MCMC proposto por Chantala e Suchindran (2005)

4.6.3 Regras de Rubin

As regras de Rubin (*Rubin rules*) estão amplamente divulgadas na literatura que trata de imputação múltipla, pois se trata de regras simples que resolvem o passo três da IM, isto é, a combinação dos resultados obtidos em diferentes análises. Essas regras podem ser usadas independentemente do método que se usou para se fazer a IM (Schafer e Graham, 2002).

A partir das **m** imputações realizadas, o passo dois 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 de um modo simples e apropriado como proposto por Rubin (1987).

A idéia é que, a partir de cada análise, obtenham-se estimativas para o parâmetro de interesse Q, ou seja, Q_j para j=1, 2,...,m. Segundo Schafer (1999), 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: $T = \overline{U} + \left(1 + \frac{1}{m}\right)B$.

4.6.4 Aplicativos que fazem imputação

Para a realização da parte computacional da imputação de dados alguns aplicativos têm sido bastante referidos na literatura para a implementação dos métodos citados. Dentre os mais utilizados pode-se citar os aplicativos SAS, S-Plus, SOLAS, NORM, BMDP e MICE, sendo que o MICE é de domínio público, pois é feito no R, fato que facilita sua utilização. Uma análise do desempenho dos aplicativos computacionais para IM pode ser vista em Horton e Lipsitz (2001) e Horton e Kleinman (2007).

Acock (2005) apresenta em seu trabalho um quadro que está reproduzido aqui, listando vários aplicativos que tem implementado em suas rotinas métodos de imputação única e/ou IM. No quadro 1 estão indicados os aplicativos, há sinalização sobre que métodos estão implementados e ainda indicação das páginas da WEB onde podem ser obtidas informações sobre os mesmos aplicativos. Como esses aplicativos estão em constantes revisões é interessante que se consulte suas páginas na web para o conhecimento das atualizações feitas. Segundo Acock (2005) devem estar contados os

dias que as revistas científicas aceitarão trabalhos sem análise adequada de dados faltantes, exceto onde abordagens tradicionais possam ser justificadas.

Quadro 1: Aplicativos usados para análise de dados faltantes

Aplicativo	Página na WEB	Imputação única	Imputação múltipla
Grátis			
Amelia	http://gking.harvard.edu/amelia/		✓
CAT	http://www.stat.psu.edu/~jls/misoftwa.html#aut	✓	
EMCOV	http://methcenter.psu.edu/downloads/EMCOV.html	✓	
NORM	http://www.stat.psu.edu/~jls/misoftwa.html#aut	✓	✓
MICE	Free with R, commercial with S-Plus http://www.multiple-imputation.com		✓
MIXED	Free with R, commercial with S-Plus http://www.stat.psu.edu/~jls/misoftwa.html#aut	✓	✓
MX	http://www.vcu.edu/mx/	✓	
PAN	Free with R, commercial with S-Plus http://www.stat.psu.edu/~jls/misoftwa.html#aut	✓	✓
Comercial			
EQS	http://www.mvsoft.com/	✓	
HLM	http://www.ssicentral.com/hlm/index.html	✓	✓
Mplus	http://www.statmodel.com	✓	✓
SAS	http://www.sas.com	✓	
SOLAS	http://www.statsol.ie/solas/imputationtechniques.htm	✓	✓
S-Plus	http://www.stat.psu.edu/~jls/misoftwa.html#aut	✓	✓
SPSS	http://www.spss.com, módulo opcional	✓	
Stata	http://www.stata.com, instalando ice ou mvis	✓	

No início do uso da IM, ou seja, na década de 80, por razões práticas, usava-se **m** pequeno, sendo que a literatura indica valores entre 3 e 10. O valor de **m** igual a 5 passou a ser o mais freqüente. Esses valores surgiram da experiência dos pesquisadores que, junto com os resultados teóricos de Rubin, viram que esses valores sugeridos para **m** são suficientes para que as conclusões sejam válidas. Entretanto, hoje em dia, com os avanços computacionais, tornou-se viável aumentar o número de imputações sem que isso cause problemas. É possível usar **m** igual a 100 ou 200 (Kenward, 2007).

O artigo de Graham et al. (2007) aborda a situação onde o número **m** é maior (**m** igual a 20, 40, 100). Em estudos de simulação os autores observaram ganho em algumas situações. É interessante que se façam mais estudos nesse sentido, pois hoje em dia com

o avanço tecnológico é possível trabalhar com **m** maiores sem que isso cause prejuízo na análise, tal como demora em se obter os resultados.

4.7 – Um roteiro simples para imputação múltipla

4.7.1 Proporção de dados faltantes

Segundo Harrell (2001) é possível serem definidas linhas gerais para a escolha entre os métodos de imputação de acordo com a proporção de dados faltantes em qualquer uma das variáveis.

(a) Proporção $\leq 0,05 \rightarrow$ Neste caso pode ser usada imputação única ou analisar somente os dados completos.

(b) Proporção entre 0,05 e 0,15 $\rightarrow$ Imputação única pode ser usada aqui provavelmente sem problemas, entretanto o uso da imputação múltipla é indicado.

(c) Proporção $\geq 0,15 \rightarrow$ A imputação múltipla é indicada na maior parte dos casos.

Obs: Se houver muitos preditores com dados faltantes devem ser feitas as mesmas considerações acima, mas os efeitos da imputação de dados serão mais pronunciados.

4.7.2 Seleção de variáveis

Como regra geral, usar toda a informação das variáveis disponíveis produz imputações múltiplas com mínimo viés e máxima precisão. Este princípio implica que o número de preditores seja tão grande quanto possível. Alguns autores observam que incluir tantos preditores quanto possível tende a fazer a suposição de MAR mais

plausível, reduzindo a necessidade de se fazer ajustes especiais para mecanismos NMAR (van Buuren e Oudshorrn, 2000).

Entretanto, conjuntos de dados frequentemente contêm muitas variáveis, todas com potencial preditivo de gerar imputações e não é factível (por problemas de multicolinearidade e computacionais) incluir todas essas variáveis. E também não é necessário. Na experiência de van Buuren, o aumento na explicação da variância na regressão linear é geralmente insignificante depois de se adicionar 15 variáveis. Para a proposta da imputação é interessante selecionar subconjuntos de covariáveis que não contenham mais do que 15 a 25 variáveis. Resumidamente, van Buuren et al. (1999) indicam a seguinte estratégia de seleção de covariáveis:

1) Incluir todas as covariáveis 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 suas correlações/associações com a resposta indicadora da variável alvo (isto é, a variável a ser imputada), ou seja, uma variável que assume o valor 1 quando o dado é faltante e o valor 0 (zero) quando o dado é observado.
3) Remover as variáveis selecionadas no passo 2 que têm muitos dados faltantes dentro do subgrupo de casos incompletos. Um indicador simples é o percentual de casos observados dentro desse subgrupo.

Usualmente, muitos preditores usados para imputação são eles mesmos, incompletos. A princípio, poderia ser aplicada a modelagem acima para cada preditor incompleto, mas isso 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 essas variáveis chave.

Em geral, os modelos de imputação devem conter variáveis com potencial poder preditivo da perda e acomodar a estrutura, por exemplo, de interações. Falha em acomodar a estrutura dos dados pode gerar viés nos resultados. 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 (Kenward et al., 2007).

Por causa da natureza Bayesiana da IM, na situação de super ajuste do modelo, isto é, incluir preditores redundantes, pode ser esperado que haja redução de precisão nas estimativas finais, mas não muita, o que não deve causar outros problemas, como viés. Em contraste a isso, omissão de importantes preditores da perda pode gerar viés. À luz disso pode ser melhor errar para mais, isto é, super ajustar do que para sub ajustar (Kenward et al., 2007).

4.8 Imputação única ou IM?

Os métodos de imputação única, que são regularmente usados na prática, provavelmente pela sua simplicidade, mostram ganho nos resultados em relação à análise de casos completos. Entretanto, esses métodos podem reduzir a variabilidade amostral por imputarem valores centrais da distribuição (método da substituição pela média ou mediana) para todos os sujeitos com dados faltantes (Little e Rubin, 2002).

Embora os métodos de imputação única não tornem os resultados viesados, superestimam a precisão, isto é, produzem erros padrão muito pequenos, enquanto a IM

produz resultados não viesados e com erros padrão apropriados (Twisk et al., 2002; Clark et al., 2003; van der Heijden et al., 2006).

Uma vantagem da IM em relação à imputação única é que a IM leva em conta a variabilidade entre imputações e, por ter o componente Bayesiano no procedimento, faz com que não aconteça 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 (Ambler et al., 2007).

É comum que estudos que tenham por objetivo encontrar modelos preditivos, o façam através de análise de regressão. Portanto, pode ser um problema usar técnicas de imputação única se o método de seleção das variáveis preditoras baseia-se nos valores-p. Como a imputação única não permite levar em conta a variabilidade entre imputações, uma conseqüência disso é que os intervalos de confiança dos coeficientes de regressão podem ser muito estreitos e os valores-p muito pequenos. Técnicas de IM têm a vantagem de levar em conta a incerteza das imputações, portanto são geralmente recomendadas para o desenvolvimento de modelos preditivos porque geram inferências corretas que refletem de maneira adequada a incerteza devido aos dados faltantes (Little e Rubin, 2002; Ambler et al., 2007).

Baseados em estudos de simulação e por embasamento teórico van der Heijden et al. (2006) defendem fortemente a idéia de que imputar dados faltantes é melhor do que ignorá-los. Essa noção, entretanto, parece que ainda não está presente entre os estudos na área de saúde porque muitos estudos ainda apresentam análises restritas aos casos completos. Ainda é pequeno o número de estudos que usam dados empíricos nos quais diferentes métodos para lidar com dados faltantes são aplicados e os resultados comparados (van der Heijden, 2006).

A partir de informações da literatura pode-se dizer que a imputação única por regressão é melhor que a imputação única da substituição pela média e que a IM é melhor que a imputação única. Também há referências que dizem que a imputação pela regressão é tão boa quanto a IM em termos de precisão, entretanto fornecem erros padrão menos realistas (Ambler et al., 2007).

Mais estudos empíricos com maiores proporções de dados faltantes devem ser feitos para mostrar como as diferenças tornam-se maiores e a favor da IM. Para ajudar os pesquisadores da área médica, mais trabalhos com foco na metodologia devem ser produzidos, indicando quando se deve usar a IM e não a imputação única (van der Heijden, 2006).

É importante para a área da Epidemiologia que mais estudos incluam resultados obtidos por técnicas de tratamento de dados faltantes. Para isso, os pesquisadores precisam estabelecer estratégias de como lidar com os dados faltantes no planejamento do processo de modelagem (Harrell, 2001; Omar et al., 2004; Ambler et al., 2007).

4.9 Algumas aplicações de imputação

O pesquisador deve analisar suas opções e, na fase de análise, certificar-se de que a imputação ou o abandono dos casos incompletos não agregou vieses relevantes aos resultados obtidos. Vários trabalhos publicados que relatam estudos de populações diversas descrevem a realização de imputação de dados.

Edwards et al. (2001), em pesquisa realizada com dados de mais de 90.000 pacientes para desenvolver modelos estatísticos de risco para mortalidade cirúrgica após cirurgias cardíacas, relatam que para todas as variáveis contínuas, os dados faltantes foram substituídos pela mediana e para as variáveis categóricas o valor imputado foi aquele mais prevalente na população.

Khuri et al. (1997) relatam os resultados de pesquisa desenvolvida com dados de mais de 87.000 pacientes para o desenvolvimento de um modelo de ajuste de risco para avaliação da qualidade dos serviços cirúrgicos dos hospitais ligados ao *National Veterans Affairs* através das taxas de mortalidade cirúrgica, sendo que foi realizado um processo de imputação de dados laboratoriais faltantes para 39% dos pacientes, pois a existência desses na base de dados dependia da solicitação da equipe médica. Um procedimento de regressão foi utilizado para estimar os valores faltantes. O valor da albumina sérica foi identificado como o principal preditor de mortalidade cirúrgica nesse estudo.

Klück (2004) apresenta um estudo feito no Hospital de Clínicas de Porto Alegre com 651 pacientes e assim como Khuri et al. (1997) realizaram a imputação de resultados laboratoriais faltantes, no caso, para a albumina sérica, pela mediana e também pelo valor normal da população. Dados comparativos do modelo de regressão logística são comparados para os dois tipos de substituição dos dados faltantes. A partir do modelo encontrado para um escore de risco multifatorial para mortalidade cirúrgica pós laparotomia exploradora, concluíram que a albumina sérica é um importante preditor e que o tipo de imputação única utilizada não altera substancialmente o modelo final.

Clark e Altman (2003) discutem em seu trabalho a importância da imputação de dados para evitar a perda de poder estatístico e a introdução de vieses de informação. Em uma amostra de 1189 casos (842 óbitos) de câncer de ovário, os casos completos eram apenas 518 (380 óbitos). Através da imputação múltipla foi construído um modelo com maior capacidade preditiva. Os autores concluíram que valores faltantes podem ser imputados em casos onde a razão para o dado estar ausente é conhecida, particularmente

quando esse pode ser explicado por variáveis conhecidas. Este procedimento aumenta o poder da análise e produz modelos mais confiáveis e aplicáveis na prática clínica.

Segundo Daley et al. (2003), dados faltantes são um problema comum em estudos clínico-epidemiológicos, principalmente quando são retrospectivos. O sistema APACHE, que avalia gravidade de pacientes admitidos a unidades de tratamento intensivo, assume o valor considerado como limite normal para todas as variáveis laboratoriais faltantes. Mas em algumas situações, a ausência de informações pode ser devida a motivos bem diferentes e esta solução não se mostrar aplicável. Principalmente em relação a resultados de exames laboratoriais, deve-se entender porque estes estão faltantes, se devido à conduta do médico, necessidade clínica ou preferências do paciente ou mesmo se o paciente saiu do estudo antes dos testes serem realizados.

No artigo de Clogg et al. (1991) os autores descrevem a metodologia usada no maior projeto do *US Bureau of the Census* que foi a recodificação do sistema de códigos de indústria e ocupação. Este projeto representou a mais extensa aplicação de imputação múltipla a dados e o empenho na modelagem também foi considerável, pois centenas de regressões logísticas foram realizadas. Um dos objetivos do artigo foi resumir as estratégias usadas no projeto para que pesquisadores entendessem como o novo banco de dados foi criado. Outro objetivo foi mostrar como modificações do método de máxima-verossimilhança foram feitas para a modelagem e fases de imputação do projeto, fazendo uso de regressão logística Bayesiana. Para muitos dos modelos de regressão, os dados eram muito dispersos para suportar as análises convencionais de máxima-verossimilhança, então modelos alternativos tiveram que ser usados. Esses métodos solucionaram problemas frequentemente encontrados com métodos de máxima-verossimilhança. A estratégia Bayesiana usada nesse projeto pode ser aplicada

em outros conjuntos de dados dispersos onde modelos logísticos são usados porque essa abordagem pode ser facilmente implementada com qualquer programa computacional.

Little (1992) faz uma revisão da literatura da análise de regressão linear com dados faltantes nas variáveis independentes. Seis classes de procedimentos são descritas: análise com casos completos (são excluídas da análise as unidades com dados faltantes); métodos de casos disponíveis (os parâmetros são estimados separadamente com o maior conjunto de unidades com dados completos disponíveis); mínimos quadrados (os dados são imputados por mínimos quadrados); máxima verossimilhança (algoritmo EM); métodos Bayesianos (amostrador de Gibbs e dados aumentados) e imputação múltipla. Os métodos são comparados e ilustrados através de um exemplo quando os dados faltantes acontecem em uma variável independente e extensões para padrões mais gerais são indicados. É dada atenção particular para quando os dados faltantes não são gerados por mecanismos completamente aleatórios. O autor sugere que é necessário mais enfoque na parte computacional do que avanços nos métodos de análise de dados completos, de casos disponíveis e imputações simples.

Muitos, senão a maioria dos estudos epidemiológicos defronta-se com o problema de dados faltantes para variáveis e/ou sujeitos. Embora a análise de dados com valores faltantes tenha sido bastante estudada na literatura estatística, os livros de epidemiologia têm dado muito pouca atenção para o assunto. Isto já era notado por Greenland e Finkle (1995) na década passada, mas ainda hoje é observado. Uma possível justificativa para isso é que os métodos que têm sido discutidos, como máxima verossimilhança, imputação múltipla e equações de estimação ponderadas, são bastante sofisticados e sua complexidade e falta de aplicativos estatísticos disponíveis dificultam sua utilização.

Tais métodos complexos têm sido discutidos porque métodos simples para se lidar com dados faltantes, como estratificação dos dados de acordo com o status de *missing*, imputação pela média condicional ou a análise com casos completos têm mostrado que fornecem resultados viesados sob determinadas circunstâncias. Greenland e Finkle (1995) revisaram esses resultados no contexto de regressão logística e apresentam estudos de simulação que mostram as limitações desses métodos. Os autores comparam os resultados da imputação múltipla com métodos simples de análise através de um estudo de caso-controle de câncer do endométrio e eles encontram uma diferença significativa nos resultados para a idade da menarca. Em geral, os autores recomendam que os epidemiologistas evitem os métodos simples e usem os métodos mais sofisticados sempre que uma grande proporção dos dados for faltante .

Barnard e Meng (1999) revisam o método de imputação múltipla de Rubin através de três aplicações que são relevantes para os estudos médicos. O artigo faz uma descrição do método de imputação múltipla e em todas as aplicações dá um forte enfoque para o primeiro passo do método de IM que é o aspecto fundamental da imputação. A primeira aplicação é sobre estimar o tempo de sobrevivência de pacientes de AIDS após o diagnóstico a partir de dados dos sistemas de vigilância. A segunda é sobre dados faltantes e não-acompanhamento em experimentos aleatorizados, onde um experimento de escolha da escola é usado como ilustração. A terceira aplicação é no estudo das não-respostas no levantamento *United States National Health and Nutrition Examination Surveys* (NHANES).

Os autores concluem que o método de Rubin é uma poderosa ferramenta e tem a grande vantagem de flexibilidade em se manusear os dados faltantes. Entretanto, é importante que se tenha cautela, assim como em qualquer metodologia estatística que se faça uso. É claro que se o modelo de imputação não consegue capturar o mecanismo de

não-resposta, as análises com tais imputações estarão comprometidas. Esse problema só pode ser evitado investigando-se cuidadosamente, em cada situação específica, o conhecimento acerca do mecanismo de não-resposta. Essa não é uma particularidade do método de Rubin, mas sim um procedimento que deve se fazer sempre que se tiverem dados incompletos. Assim como em qualquer método estatístico, a abordagem de Rubin para imputação múltipla não tem uma receita universal para todos os problemas de dados faltantes.

van Buuren et al. (1999) estudaram o problema de não-resposta na análise de sobrevivência onde a ocorrência de não-resposta nos fatores de risco está relacionada com a mortalidade. Num estudo para determinar a influência da pressão sanguínea na sobrevivência de idosos (85 anos ou mais), as medidas de pressão foram perdidas em 12,5% da amostra. Os dados disponíveis sugerem que o processo que gerou a não-resposta depende conjuntamente da sobrevivência e do valor desconhecido da pressão, implicando na distorção da relação de interesse. A imputação múltipla é usada para imputar a pressão sanguínea e então analisar os dados sob uma variedade de modelos de não-respostas. Uma modelagem é tratada em detalhes: a construção de um modelo preditivo para gerar imputações se o número de covariáveis é grande. Estimativas dos riscos para esses dados parecem robustas mesmo quando grandes afastamentos do modelo mais simples de não-resposta acontecem, e são similares àquelas obtidas quando não se utilizaram os casos incompletos.

O artigo de Zhou et al. (2001) mostra resultados a partir de dois estudos reais, o primeiro com dados faltantes no desfecho (estudo da satisfação dos pacientes com a troca de médicos por causa do fim da residência) e o segundo com dados faltantes nas covariáveis (resultados de exames laboratoriais num estudo de mortalidade por asma) e mostra também um estudo de simulação para avaliar a performance do método de

imputação. Os autores ficaram surpresos com o fato de que a imputação múltipla apresentou resultados similares às imputações únicas de média e mediana feitas para os dois estudos. A melhor explicação que eles encontraram para isso foi o fato de estarem lidando com um número grande de observações. Os autores concluem que pode ser difícil haver muita diferença entre as estimativas finais dos diferentes métodos de imputação, e, portanto, as conclusões de um grande estudo com diferentes métodos de imputação não são, em geral muito afetadas pela escolha do método, a não ser que os resultados estejam no limiar da significância. Num dos estudos eles mostram que somente a análise com dados completos poderia dar resultados diferentes da análise com dados imputados. No entanto, a simulação mostrou melhores estimativas para os erros padrão na imputação múltipla do que a imputação única pela média predita.

O trabalho de van der Heijden et al. (2006) usou dados de 398 pacientes com 18 anos ou mais que procuraram um hospital por suspeita de embolia pulmonar. Os autores avaliaram quais resultados dos testes diagnósticos (preditores) contribuíram para predizer a presença ou ausência da embolia e utilizaram cinco diferentes métodos para lidar com os dados faltantes: análise somente com os casos completos, três métodos de imputação única e IM. O objetivo principal dos autores era mostrar o efeito das cinco maneiras diferentes de se trabalhar com dados faltantes. Para uma ampla visão técnica dos métodos, eles indicam a literatura: Schafer, 2002; Rubin, 1976; Greenland e Finkle, 1995; Vach, 1997; Little, 1992; Schafer, 1997; Clark e Altman, 2003; Harrell, 2001 e van Buuren et al., 1999. As imputações únicas foram feitas pelo SPSS e para a IM foi usado o MICE.

As medidas utilizadas para comparar os métodos dos modelos foram a direção e magnitude dos coeficientes de regressão, os erros padrão e as curvas ROC (Hanley, 1982). Os autores encontraram diferenças consideráveis no modelo obtido com a análise

de casos completos – abordagem muito comum em estudos epidemiológicos, principalmente estudos de diagnósticos – em relação aos modelos obtidos com os dados imputados no que se refere aos preditores selecionados, coeficientes de regressão e correspondentes erros padrão. Não houve muita diferença entre os modelos obtidos com os diferentes métodos de imputação. Também é citado que não houve muita diferença nos resultados obtidos pela imputação única da regressão e a IM. Os autores justificam isso pelo fato dos preditores terem poucos dados faltantes.

Moons et al. (2006) usaram dados de um estudo de embolia pulmonar, em que foram selecionados cinco preditores de embolia que não tinham dados faltantes. Os coeficientes de regressão e os erros padrão obtidos desse conjunto de dados foram considerados como valores verdadeiros. A partir de um processo de simulação foram obtidas amostras que tinham dados faltantes nos preditores, sendo que os autores simularam amostras sob as suposições MCAR e MAR. Os dados faltantes foram imputados por IM considerando as situações com ou sem o desfecho nos modelos da IM. Nas conclusões os autores comentam que os coeficientes obtidos com os dados imputados pela IM que incluiu o desfecho foram mais próximos dos coeficientes reais. A IM sem o desfecho produziu coeficientes subestimados. Os resultados foram os mesmo para os dados MCAR e MAR.

Shrive et al. (2006) apresentam uma aplicação de imputação a dados de saúde mental. Através de um estudo que usou uma escala de depressão foram comparados vários métodos de imputação. A amostra foi composta de 1580 pacientes participantes de um estudo de desfechos cirúrgicos. Os pacientes preencheram um questionário de 20 questões usado para medir depressão, o instrumento *Zung Self-reported Depression Scale* (SDS). Todas as questões do questionário são preenchidas com valores entre 1 e 4. O escore do paciente é a soma de todas os ítens e escore acima de 40 indica presença

de sintomas depressivos. Dados faltantes foram simulados aleatoriamente de forma que foram gerados três amostras com dados faltantes: MCAR, MAR e NMAR. Foram usados seis métodos de imputação: 1) IM; 2) imputação única pela regressão; 3) imputação única pela média individual; 4) imputação única pela média geral; 5) resposta precedente do indivíduo e 6) escolha aleatória entre os valores 1 e 4. Para cada método foram calculadas as medidas de escore médio e desvio padrão que foram usadas para a comparação dos métodos. A IM foi o melhor método para lidar com os dados faltantes na maioria dos cenários criados. Entretanto, os autores concluem que o método da imputação única pela média individual também foi apropriado para lidar com os dados faltantes em alguns cenários de dados faltantes e pode ser mais facilmente interpretado pelos leitores médicos.

O trabalho de Arnold e Kronmal (2003) refere-se ao *Cardiovascular Health Study* que tem por objetivo identificar fatores de risco para doenças cardiovasculares em indivíduos de 65 anos ou mais. Foi ajustado um modelo preditivo para enfarte que incluiu variáveis preditoras que foram medidas no início do estudo entre pacientes sem história prévia de enfarte. Tal trabalho descreve a dificuldade de descrever o processo de se imputar mais que 100 variáveis e comparar os resultados obtidos com a análise restrita aos dados completos e os resultados obtidos com os dados imputados. Os autores usaram o S-Plus. Foram ajustados modelos de Cox e as medidas de comparação foram as razões de risco (*hazard ratios*) e respectivos intervalos de confiança. Os resultados obtidos com a análise de dados completos e a análise com dados imputados foram similares. Ainda, os resultados da imputação única foram muito pouco diferentes da IM. Entretanto, os autores não se surpreenderam com os resultados, pois aproximadamente 85% das variáveis com dados faltantes tinham no máximo 5% de perda, implicando que a quantidade de dados imputados foi muito pequena. Apesar dos

resultados obtidos, os autores salientam a vantagem de se utilizar imputação, dado que a amostra não perde em poder. Entretanto, também comentam que nesse estudo o tamanho da amostra não é problema (n=5002).

4.10 Uma aplicação para modelos de risco

No trabalho de Ambler et al. (2007) que trata do desenvolvimento de um modelo de risco para mortalidade hospitalar foram avaliados três métodos de imputação única (média, média/moda e média condicional) e dois métodos de IM. O trabalho teve como foco desfechos binários, comumente usados em desenvolvimento de modelos de risco. Muitos modelos de risco têm sido desenvolvidos para esse tipo de desfecho e frequentemente são usados para monitorar o desempenho de instituições hospitalares. No contexto de variáveis binárias é natural que se use modelos de regressão logística para modelar o log(odds) do desfecho como uma combinação linear dos preditores. Especificamente, nesse trabalho, os autores usaram como desfecho a mortalidade hospitalar de pacientes submetidos à cirurgia cardíaca a partir do conjunto de dados da *Society of Cardiothoracic Surgeons of Great Britain and Irelan*.

Os métodos de imputação utilizados no trabalho de Ambler et al. (2007) consideraram a suposição de que os dados faltantes são MCAR ou MAR. Foi feita a avaliação de alguns métodos de imputação que comumente são usados no contexto de desenvolvimento de modelos de risco. Está claro que os dados faltantes podem afetar a predição dos modelos de risco e simplesmente ignorar os dados faltantes e analisar os dados completos pode levar a viés nos resultados ou empobrecimento da predição, o que, na prática, afetaria as estratégias de tratamento e decisões, acarretando no caso de mortalidade cirúrgica, sérias implicações clínicas.

Os autores alertam para o fato de que se deve ter cuidado para a generalização dos resultados obtidos, pois eles não podem ser generalizados para qualquer situação. Mais trabalhos são necessários para avaliar o desempenho de métodos de imputação para conjuntos de dados com maior proporção de preditores contínuos com possível relação não linear e associação mais forte entre os preditores, bem como conjuntos de dados com desfechos não binários. Os autores usaram o Stata, entretanto citam que a análise pode também ser realizada pelo MICE.

Como conclusão, é recomendado que os pesquisadores ao estimarem modelos de risco clínico não ignorem simplesmente o problema de dados faltantes fazendo a análise somente com os casos completos, como freqüentemente é feito. Imputar dados faltantes pode aumentar consideravelmente a confiabilidade das predições de risco obtidas com os modelos ajustados, como mostrado no trabalho. Além disso, estratégias para se lidar com dados faltantes podem aumentar o tamanho efetivo do conjunto de dados e permitir a inclusão de mais preditores no modelo, como já foi citado por outros autores (por exemplo,Harrell, 2001).

Métodos de imputação única podem ser apropriados, desde que a estratégia de seleção de variáveis para o modelo de risco não seja baseada nos valores-p do ajuste, visto que a imputação única fornece boa precisão, entretanto subestima a variabilidade. Métodos de IM são geralmente recomendados para o desenvolvimento de modelos de risco porque eles geram inferências corretas que refletem de maneira adequada a incerteza devida aos dados faltantes (Little, 2002).

5. REFERÊNCIAS BIBLIOGRÁFICAS

(1) Acock, AC. Working with missing values. **Journal of Marriage and Family**, 2005; 67: 1012–1028.

(2) Ambler G, Omar RZ, Royston P. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. **Statistical Methods in Medical Research**, 2007; 16(3): 277-298.

(3) Arnold, AM e Kronmal, RA. Multiple imputation of baseline data in the Cardiovascular Health Study. **American Journal of Epidemiology**, 2003; 157(1): 74-84.

(4) Barnard J and Meng XL. Applications of multiple imputation in medical studies: from AIDS to NHANES. **Statistical Methods in Medical Research**, 1999; 8:17 –36.

(5) Box, GEP and Tiao, GC. **Bayesian inference in statistical analisys**. Reading, Mass: Addison-Wesley, 1973.

(6) Chantala, K and Suchindran, C. **Multiple imputation for missing data.** Em: http://www:cpc:unc:edu=services=computer=presentations=mipresentation2:pdf, aces-so em 15/11/2005.

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

(8) Clogg CC, Rubin DB, Schenker N, Schultz B, Weidman L. Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression (in applications and case studies). **Journal of the American Statistical Association**, 1991; 86:68 –78.

(9) Daley, J; Iezzoni, LI and Shwartz, M. Conceptual and practical issues in developing risk adjusted methods. In: Iezzoni LI, editor. Risk adjustment for measuring health care outcomes. Chicago, Illinois: **Health Administration Press**, 2003; 179-205.

(10) Dempster, AP, Laird, NM & Rubin, DB.. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). **Journal of the Royal Statistical Society, Series B**, 1977; 39:1–38.

(11) Diggle, PJ, e Kenward, MG. Informative dropout in longitudinal data analysis (with discussion). **Applied Statistics**, 1994; 43:49–73.

(12) Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. **Journal of Clinical Epidemiology,** 2006 59(10): 1087-1091.

(13) Edwards, FH; Peterson, ED; Coombs, LP; Delong, ER; Jamieson, WR; Shroyer, ALW et al.. Prediction of operative mortality after valve replacement surgery. **Journal of the American College of Cardiology**, 2001; 37(3):885-892.

(14) Engels JM, Diehr P. Imputation of missing longitudinal data: a comparison of methods, **Journal of Clinical Epidemiology**, 2003; 56(10):968-76.

(15) Fraser G e Yan R. Guided multiple imputation of missing data: using a subsample to strengthen the missing-at-random assumption. **Epidemiology**, 2007; 18(2): 246-52.

(16) Gilks, WR; Richardson, S e Spiegelhalter, DJ. **Markov chain Monte Carlo in practice**. London:.Chapman & Hall, 1996.

(17) Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. **Prevention Science**, 2007; 8(3): 206-213.

(18) Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. **American Journal of Epidemiology**, 1995; 142 (12):1255-1264.

(19) Hanley JA e McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. **Radiology**, 1982; 143: 29-36.

(20) Harel O, Zhou XH. Multiple imputation: review of theory, implementation and software. **Statistics in Medicine**, 2007; 26(16): 3057-77.

(21) Harrell, FE Jr. **Regression modeling strategies with applications to linear models, logistic regression and survival analysis.** Springer–Verlag, New York, 2001.

(22) Horton, NJ and Lipsitz, SR. Multiple imputation in practice: Comparison of software packages for regression models with missing variables. **The American Statistician,** 2001; 55(3):244-54.

(23) Horton, NJ e Kleinman, KP. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. **The American Statistician**, 2007, 61(1): 79-90.

(24) Kenward MG e Carpenter J. Multiple imputation: current perspectives. **Statistical Methods in Medical Research**, 2007; 16(3): 199-218.

(25) Khuri, SF; Daley, J; Henderson, W; Hur, K; Gibbs, JO; Barbour, G et al.. Risk adjustment of the postoperative mortality rate for the comparative assessment of the quality of surgical care: results of the National Veterans Affairs Surgical Risk Study. **Journal of the American College of Surgeons**, 1997; 185(4):315-327.

(26) Klück, M. **Metodologia para ajuste de indicadores de desfechos hospitalares por risco prévio do paciente**. Tese de doutorado em Epidemiologia. Faculdade de Medicina. Universidade Federal do Rio Grande do Sul .Porto Alegre, 2004.

(27) Lavori P, Dawson R e Shera D. A multiple imputation strategy for clinical trials with truncation of patient data. **Statistics in Medicine,** 1995; 14:1913-25.

(28) Little, RJA. Regression with Missing Xs - A Review. **Journal of the American Statistical Association**, 1992; 87(420):1227-37.

(29) Little, RJA e Rubin, DB. **Statistical analysis with missing data**. New York: (Second Edition) Wiley, 2002.

(30) McCullagh, P e Nelder, JA. **Generalized linear models**. London: Chapman & Hall, 1989.

(31) Moons KG, Donders RA, Stijnen T, Harrell FE Jr. Using the outcome for imputation of missing predictor values was preferred. **Journal of Clinical Epidemiology**, 2006; 59(10): 1092-1101.

(32) Neyman, J e Pearson, ES. On the problem of most efficient tests of statistical hypotheses. **Philosophical Transactions of the Royal Society of London, Series A**, 1933; 231:289–337.

(33) Neyman, J. Outline of a theory of statistical estimation based on the classical theory of probability. **Philosophical Transactions of the Royal Society of London, Series A**, 1937; 236:333–380.

(34) Omar RZ, Ambler G, Royston P, Eliahoo J, Taylor KM. Cardiac surgery risk modeling for mortality: a review of current practice and suggestions for improvement. **Annals of Thoracic Surgery**, 2004; 77: 2232–2237.

(35) Rosenbaum, PR e Rubin, DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. **Biometrika**, 1983; 70:41-55.

(36) Rubin, DB. Inference and missing data. **Biometrika**, 1976; 63:581–592.

(37) Rubin, DB. **Multiple Imputation for Nonresponse in Surveys.** New York: Wiley, 1987.

(38) Rubin, DB. Multiple imputation after 18+ years. **Journal of the American Statistical Association**, 1996; 91(434):473-89.

(39) Rubin, DB e Schenker, N. Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. **Journal of the American Statistical Association**, 1986; 81:366-374.

(40) Schafer JL. **Analysis of incomplete multivariate data**. London: Chapman & Hall/CRC Press; 1997.

(41) Schafer, JL. Multiple imputation: a primer. **Statistical Methods in Medical Research**, 1999; 8(1):3-15.

(42) Schafer, JL, and Graham, JW. Missing data: our view of the state of the art. **Psychological Methods**, 2002; 7:147-177.

(43) Shrive, FM; Stuart, H; Quan H e Ghali, WA. Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. **BMC Medical Research Methodology**; 2006, 6:57. Este artigo está disponível em http://www.biomedcentral.com/1471-2288/6/57

(44) Tang L, Song J, Belin TR, Unützer J. A comparison of imputation methods in a longitudinal randomized clinical trial. **Statistics in Medicine**, 2005; 24: 2111-28.

(45) Twisk J e de Vente W. Attrition in longitudinal studies: How to deal with missing data. **Journal of Clinical Epidemiology**, 2002; 55: 329-337.

(46) Vach W. Some issues in estimating the effect of prognostic factors from incomplete covariate data. **Statistics in Medicine,** 1997; 16: 57-72.

(47) van Buuren, S and Oudshoorn CGM. **Multivariate imputation by chained equations. MICE V1.0 User's Manual**. Report PG/VGZ/00.038. Leiden: TNO Preventie en Gezondheid, 2000.

(48) van Buuren, S, Boshuizen, HC and Knook, DL. Multiple imputation of missing blood pressure covariates in survival analysis. **Statistics in Medicine**, 1999; 18:681-694.

(49) 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. **Journal of Clinical Epidemiology**, 2006; 59(10): 1102-1109.

(50) White, IA; Wood, A e Royston, P. Editorial: Multiple imputation in practice. **Statistical Methods in Medical Research**, 2007; 16; 195-197.

(51) Zhang, P. Multiple imputation: Theory and method. **International Statistical Review**, 2003; 71(3):581-92.

(52) Zhou, XH, Eckert, GJ and Tierney, WM. Multiple imputation in public health research. **Statistics in Medicine**, 2001; 20(9-10):1541-49.

ARTIGO 1

GANHO DE PRECISÃO COM A IMPUTAÇÃO MÚLTIPLA EM RELAÇÃO À EXCLUSÃO DE CASOS COM OBSERVAÇÕES FALTANTES

PROFIT OF PRECISION WITH THE MULTIPLE IMPUTATION IN RELATION TO THE EXCLUSION OF CASES WITH MISSING DATA

Luciana Neves Nunes, ***doutoranda em Epidemiologia pela UFRGS.***

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL (UFRGS)

A ser enviado aos Cadernos de Saúde Pública

GANHO DE PRECISÃO COM A IMPUTAÇÃO MÚLTIPLA EM RELAÇÃO À EXCLUSÃO DE CASOS COM OBSERVAÇÕES FALTANTES

PROFIT OF PRECISION WITH THE MULTIPLE IMPUTATION IN RELATION TO THE EXCLUSION OF CASES WITH MISSING DATA

Luciana Neves Nunes[1,2] , Mariza Machado Klück[1,3] e Jandyra Maria Guimarães Fachel[1,2]

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

2- Departamento de Estatística, Instituto de Matemática,Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil.

3- Departamento de Medicina Social, Faculdade de Medicina, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil

Endereço para correspondência

Luciana Neves Nunes (lununes@mat.ufrgs.br)

Universidade Federal do Rio Grande do Sul

Instituto de Matemática - Departamento de Estatística

Av. Bento Gonçalves, 9500

Bairro Agronomia

Porto Alegre – RS

CEP: 91509-900

RESUMO

Introdução: Em situações com dados faltantes, uma abordagem bastante comum é restringir a análise aos sujeitos com dados completos nas variáveis. Porém, as estimativas obtidas com tais análises podem ser viesadas, especialmente se os indivíduos que permanecem na análise são diferentes daqueles que foram excluídos. "Preencher" os dados faltantes é a chamada técnica de imputação. Esse trabalho tem como objetivo divulgar o método de imputação múltipla.

Método: A partir de um conjunto de dados de 470 pacientes cirúrgicos, foram ajustados modelos de regressão logística para o desfecho óbito. Todos os pacientes tinham informações completas. Por simulação foram gerados dois conjuntos de dados incompletos, um com 5% e outro com 20% de dados faltantes para uma variável. Os modelos logísticos foram ajustados em três diferentes situações: para o conjunto completo, para os conjuntos com dados faltantes e para o conjunto completado por imputação múltipla.

Resultados: As estimativas obtidas pela análise dos conjuntos com dados faltantes foram diferentes das estimativas obtidas com o conjunto completo, principalmente as do conjunto com 20% de dados faltantes. Os resultados obtidos com o banco completado por imputações foram próximos dos valores obtidos com o conjunto completo, porém um coeficiente deixou de ser estatisticamente significativo.

Conclusões: Os resultados indicam que a imputação múltipla utilizada parece ter sido eficiente, pois as estimativas tanto para o conjunto completo como para o conjunto com dados imputados foram muito similares. A imputação múltipla se mostrou superior à análise do conjunto com dados faltantes, que desconsiderou os casos incompletos, como tratamento a dados faltantes.

Palavras-chave: Imputação múltipla; não-resposta; dados faltantes; método de imputação.

ABSTRACT

Introduction: Usually in survey research subjects with nonresponse are not considered for the statistical analysis. Nevertheless, especially if the subjects considered in the analysis are different of those who have been excluded, the estimates can be biased. The method to “fill in” the missing data is called the imputation method. This work has for objective to divulge a multiple imputation method.

Method: From a data set of 470 surgical patients different logistic regression models were developed for the oucome death. All patients had complete informations. By simulation two incomplete data sets were generated, a data set with 5% and another with 20% of missing data in a single variable. The logistic models had been adjusted in three different situations: for the complete data set, for the incomplete data sets and for the data set completed by multiple imputation.

Results: The estimates obtained by the analyses of the data set with missing data were different from those observed in the complete data set, mainly the estimates of the data set with 20% of missing data. When considered the imputed data, the logistic models had estimates closer to those of the complete data set, however a coefficient became not significant.

Conclusions: The results indicate that the used multiple imputation methods used are efficient, producing results very similar to the results obtained with the complete data

set. The analysis using multiple imputation was superior to the analysis of the incomplete data sets that excluded incomplete cases in the analysis.

Key-words: Multiple imputation; nonresponse; missing data; imputation method.

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 predizer 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 e 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 esse 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 através 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 IM 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 (01/09/2007), 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 sua edições a artigos sobre tratamentos a dados faltantes e principalmente à IM. Estes fatos revelam que o estudo de metodologias para dados faltantes vem sendo bastante debatido mais recentemente, o que indica a pertinência desse 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 sua introdução há mais ou menos 30 anos atrás, 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 desse 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 ao invés de fazer a análise restrita aos casos completos. A comparação dos resultados com os dados imputados por IM e da análise restrita a casos completos será feita através de um conjunto de dados reais.

Método

Imputação Múltipla

D. B. Rubin, ainda nos anos 70, propôs a técnica de *imputação múltipla* (IM) para resolver o problema de não-resposta em pesquisas. No entanto, apenas recentemente esta 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 de três passos:

1. São obtidos ***m*** bancos de dados completos através 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 que 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[2,10].

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 as regras propostas por Rubin.[2]

As regras de Rubin (*Rubin rules*) estão amplamente divulgadas na literatura que trata de imputação múltipla, pois são normas simples que resolvem o passo três da IM. Estas regras podem ser usadas independentemente do método utilizado para fazer a IM[10]. 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 imputação múltipla. 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[11]. Análises do desempenho dos aplicativos computacionais para IM tem sido publicadas na literatura[12,13,14].

Segundo Harrell[15] 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 $\leq 0,05 \rightarrow$ Neste caso pode ser usada a imputação única ou analisar somente os dados completos.

b. Proporção entre 0,05 e 0,15 $\rightarrow$ Imputação única pode ser usada provavelmente sem problemas, entretanto o uso da imputação múltipla é indicado.

c. Proporção $\geq 0,15 \rightarrow$ A imputação múltipla é indicada na maior parte dos modelos.

Se houver muitos preditores com dados faltantes devem ser feitas as mesmas considerações acima, mas os efeitos das imputações serão mais pronunciados.

Fonte de dados

Esse artigo utilizará como exemplo um banco de dados cedido por Klück[16], trabalho em que foi desenvolvido e validado um escore de risco multifatorial para mortalidade cirúrgica pós laparotomia exploradora. A população de pesquisa foi composta de pacientes internados no HCPA de Porto Alegre, RS, no período 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 era composto de 651 pacientes, entretanto para ilustrar os métodos de imputação múltipla nesse 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 Kluck[16]. 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 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 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 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. Ou seja, os resultados obtidos com a análise do banco completo foram considerados valores verdadeiros.

A partir desse 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, respectivamente. 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) e o banco de dados com 20% de observações faltantes será referido como Banco Incompleto 20 (BI-20).

O modelo de imputação múltipla 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, dado as observadas, e então se obtém o banco completo. Esse processo é repetido **m** vezes, resultando **m** bancos completos. Nesse 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 ($Y= \alpha+\beta\mathbf{X}$), $Y \sim N(X\beta; I\sigma^2)$, 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]:

1. *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, então utiliza-se o valor observado como valor a ser imputado. A variabilidade entre imputações é gerada através dos passos que servem para estimar β e σ e que são repetidos **m** vezes.
2. Bayesian Linear Regression (*BLR*) [2]: Assim como o método *PMM*, são estimados β e σ, entretanto 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 $\geq$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 $\geq$75)+β_4(Cirurgia urgente)+Constante

A IM foi realizada através do pacote Multivariate Imputation by Chained Equatoins (MICE)[18] do programa R. Maiores detalhes computacionais podem ser vistos no Apêndice do artigo.

Quanto à inclusão de variáveis no modelo de imputação, van Buuren et al.[17] sugerem como regra geral, usar toda a informação das variáveis disponíveis, o que produz imputações múltiplas com mínimo viés e máxima precisão. Este 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[17,18].

Segundo van Buuren et al.[17], pode-se usar a seguinte estratégia de seleção de covariáveis:

1) Incluir todas as covariáveis 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 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 têm muitos dados faltantes dentro do subgrupo de casos incompletos. Um indicador simples é o percentual de casos observados dentro desse subgrupo.

Usualmente, muitos preditores usados para imputação são eles mesmos incompletos. A princípio, poderia ser aplicada a modelagem acima para cada preditor incompleto, mas isso 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 essas variáveis-chave.

Após a realização da IM com **m** igual a 5 imputações, opção *default* do MICE[18], 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[16]. As estimativas gerais foram obtidas pela aplicação das Regras de Rubin citadas anteriormente e implementadas em uma planilha do Excel.

Resultados

A Tabela 1 mostra os resultados obtidos 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 os 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 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.

Tabela 1 – Estimativas da regressão logística para o banco de dados completo e bancos incompletos (BI-5 e BI-20). Modelos ajustados com 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)
Alb 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)	(0,420)
Alb 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)

Na Tabela 2 são apresentados os resultados da regressão logística utilizando os valores imputados pelo método de imputação múltipla para a albumina, para o banco incompleto com 5% de dados faltantes (BI-5), utilizando 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.

Tabela 2 – Estimativas da regressão logística após imputações múltiplas pelo método *PMM* em diferentes regressões. Mecanismo MCAR, n=440 (BI-5).

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)
Alb 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)
Alb 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)	(0,360)

*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

A Tabela 3 mostra os resultados obtidos para a Regressão Logística com a variável albumina imputada através do método *BLR,* com diferentes configurações do modelo de regressão linear para obtenção do valor predito, para o banco de dados BI-5. È 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 usou 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 IC´s 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, quando 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.

Tabela 3 – Estimativas da regressão logística após imputações múltiplas pelo método *BLR* em diferentes regressões. Mecanismo MCAR, n=440 (BI-5).

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)	(0,350)	(0,351)	(0,351)
Alb 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)	(0,360)	(0,362)	(0,363)
Alb 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

Os resultados apresentados na Tabela 4 são os obtidos com a variável albumina imputada através do método PMM no BI-20. Os IC´s 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 a Tabela 2 e a Tabela 3, que consideram os resultados da regressão logística para dados de albumina imputados pelo método de imputação múltipla para o banco com simulação de 5% dos dados faltantes, pode-se fazer uma comparação dos resultados obtidos pelos dois métodos de IM utilizados nesse trabalho (PMM e BLR). As estimativas pontuais das RC e os respectivos IC´s foram, em geral muito parecidas 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.

Tabela 4 – Estimativas da regressão logística após imputações múltiplas pelo método *PMM* em diferentes regressões. Mecanismo MCAR, n=383 (BI-20).

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]	3,3[1,4;7,5]	3,3[1,4;7,7]	3,3[1,4;7,5]
	(0,422)	(0,423)	(0,427)	(0,423)
ASA IV/V	20,2[8,8;46,0]	20,5[8,9;47,2]	20,8[9,0;48,2]	20,5[8,9;47,2]
	(0,421)	(0,425)	(0,428)	(0,426)
Idade ≥ 75	2,9[1,5;5,8]	3,1[1,5;6,1]	3,1[1,5;6,1]	3,1[1,6;6,1]
	(0,348)	(0,350)	(0,351)	(0,349)
Alb até 2,2g/dl	5,3[2,7;10,5]	4,6[2,0;10,4]	4,5[1,8;11,3]	4,2[2,0;8,9]
	(0,349)	(0,417)	(0,466)	(0,376)
Alb 2,3 a 3,0g/dl	2,1[1,1;4,1]	1,9[0,8;4,4]	2,1[0,9;4,8]	1,8[0,8;3,9]
	(0,345)	(0,432)	(0,421)	(0,393)

*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

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 IC´s percebe-se queos 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.

Tabela 5 – Estimativas da regressão logística após imputações múltiplas pelo método *BLR* em diferentes regressões. Mecanismo MCAR, n=383 (BI-20).

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]	3,1[1,4;7,2]	3,3[1,4;7,6]	3,2[1,4;7,3]
	(0,422)	(0,427)	(0,426)	(0,424)
ASA IV/V	20,2[8,8;46,0]	19,7[8,5;45,2]	20,8[9,0;48,1]	20,0[8,7;46,1]
	(0,421)	(0,425)	(0,426)	(0,425)
Idade ≥ 75	2,9[1,5;5,8]	3,2[1,6;6,4]	3,1[1,6;6,3]	3,1[1,6;6,4]
	(0,348)	(0,351)	(0,352)	(0,358)
Alb até 2,2g/dl	5,3[2,7;10,5]	5,0[2,3;11,0]	4,5[1,9;10,7]	4,6[2,1;10,1]
	(0,349)	(0,400)	(0,447)	(0,400)
Alb 2,3 a 3,0g/dl	2,1[1,1;4,1]	2,0[1,0;4,2]	1,9[0,9;4,1]	2,1[0,9;4,8]
	(0,345)	(0,376)	(0,388)	(0,411)

*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

Ao se comparar a Tabela 4 e a Tabela 5, que consideram os resultados obtidos nos modelos de regressão logística com 20% dos dados faltantes de albumina imputados, é possível comparar os resultados obtidos pelos dois métodos de IM utilizados (PMM e BLR). O modelo obtido com a IM(1) pelo método BLR foi um pouco diferente dos demais modelos, 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 da categoria ASA III, ASA IV/Ve 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 desse 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, isto é, sem dados faltantes, com os resultados da análise feita nos bancos nos quais foram simulados faltas de dados e estes indivíduos retirados, houve discrepâncias nos valores das estimativas e aumento no tamanho dos IC´s nos bancos BI-5 e BI-20, devido a redução do tamanho amostral. Portanto, restringir a análise ao conjunto de casos que têm observações completas pode levar a conclusões erradas.[2,6,15,19]

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

Freqüentemente, em estudos epidemiológicos, modelos são estimados através de 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 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[15,19].

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 das estimações pontual e por intervalo para a RC, vê-

se que a associação dessa variável com o desfecho foi bastante fraca, pois RC=2,1 e IC95%=[1,1;4,1] e, além disso, 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, isso 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 essas alternativas implementadas computacionalmente[22].

Para a análise de regressão logística como feita nesse 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 77,0% 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.

Nesse 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 depende são bem identificadas, melhora o desempenho da IM, pois essas 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[8,9]. Entretanto, pode ser citado o trabalho de Moons et al.[23] 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[23].

Observando-se os resultados desse 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 nesse artigo[17]. 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 ajustando do que sub ajustando[8].

A partir desse único trabalho não é possível se tirar 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 isso 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. Entretanto nos últimos anos, vários aplicativos, livres e comerciais, implementaram em suas rotinas técnicas de imputação, tanto imputação única como IM. Literatura recente traz discussões acerca dessas implementações[12,13,14].

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 esse 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. Entretanto, hoje em dia, com os avanços computacionais, tornou-se praticável que o número de **m** de imputações seja muito maior sem que isso cause problemas. É possível que se use **m** igual a 100 ou 200[8,24].

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[15].

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 se deve usar técnicas de imputação para tratar o problema de dados faltantes, e ressaltando as vantagens da IM sobre a imputação única[6].

Referências bibliográficas

1. Rubin DB. Multiple imputation after 18+ years. **Journal of the American Statistical Association** 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. **Journal of the American Statistical Association** 1992; 87(420): 227-37.
4. Schafer JL. Multiple imputation: a primer. **Statistical Methods in Medical Research** 1999; 8(1): 3-15.
5. Zhang P. Multiple imputation: Theory and method. **International Statistical Review** 2003; 71(3): 581-92.
6. van der 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. **Journal of Clinical Epidemiology**, 2006; 59(10): 1102-1109.

7. White, IA; Wood, A e Royston, P. Editorial: Multiple imputation in practice. **Statistical Methods in Medical Research**, 2007; 16; 195-197.

8. Kenward MG e Carpenter J. Multiple imputation: current perspectives. **Statistical Methods in Medical Research**, 2007; 16(3): 199-218.

9. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. **Journal of Clinical Epidemiology,** 2006 59(10): 1087-1091.

10. Schafer, JL and Graham JW. Missing data: our view of the state of the art. **Psychological Methods** 2002; 7 (2): 147-177.

11. R Development Core Team R. A language and environment for statistical computing. Vienna, Austria: R **Foundation for Statistical Computing**; 2004. Available at http://www.R-project.org ISBN 3-900051-00-3, Accessed September 02, 2007.

12. Horton NJ and Lipsitz SR. Multiple imputation in practice: Comparison of software packages for regression models with missing variables. **American Statistician** 2001; 55(3): 244-54.

13. Acock, AC. Working with missing values. **Journal of Marriage and Family**, 2005; 67: 1012–1028.

14. Horton, NJ e Kleinman, KP. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. **The American Statistician**, 2007, 61(1): 79-90.

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

16. Klück M. **Metodologia para ajuste de indicadores de desfechos hospitalares por risco prévio do paciente**. Tese de doutorado em Medicina: Epidemiologia. Faculdade de Medicina. Universidade Federal do Rio Grande do Sul . Porto Alegre , 2004.

17. Van Buuren S, Boshuizen HC and Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. **Statistics in Medicine** 1999; 18:681-694.

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

19. Ambler G, Omar RZ, Royston P. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. **Statistical Methods in Medical Research**, 2007; 16(3): 277-298.

20. Schafer JL. **Analysis of incomplete multivariate data**. London: Chapman & Hall/CRC Press; 1997.

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

22. Meng, X-L. Missing data: dial M for ??? **Journal of the American Statistical Association**, 2000; 95(452): 1325-1330.

23. Moons KG, Donders RA, Stijnen T, Harrell FE Jr. Using the outcome for imputation of missing predictor values was preferred. **Journal of Clinical Epidemiology**, 2006; 59(10): 1092-1101.

24. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. **Prevention science**, 2007; 8(3): 206-213.

APÊNDICE DO ARTIGO 1

Para a realização dessas imputações como expostas acima, as instruções para o uso do o 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. Para fazer a imputação múltipla utiliza-se o pacote MICE[18] que pode ser obtido no espelho brasileiro. Os comandos apresentados mostram como foi feita a imputação múltipla pelo método PMM para o BI-20.

A primeira linha indica que o pacote MICE foi ativado. Logo a seguir foi feita a leitura do banco de dados que era do tipo *.sav*, ou seja, do SPSS, sendo a ele atribuído o nome `imputacao`. A sintaxe do aplicativo fica mais fácil se o banco de dados incluir somente as variáveis de interesse. No exemplo apresentado o banco continha as variáveis albumina, ASA, idade e caráter da cirurgia. O banco pode conter outras variáveis, entretanto, nessa situação deve-se usar um comando que explicita quais variáveis serão usadas para a imputação. Maiores detalhes podem ser obtidos no manual do MICE[18]. Depois de lido o arquivo, é interessante verificar se as variáveis foram lidas corretamente através do comando `summary`. O comando `md.pattern(nome do banco)` fornece informações sobre os dados faltantes nas variáveis. A imputação múltipla propriamente dita é feita pelo comando `mice(nome do banco)` e por

default é feita a regressão PMM. Para verificar alguns detalhes da imputação, tais como número de imputações múltiplas (***m***), quantos valores foram imputados, método de imputação e quais variáveis que tiveram dados imputados pode-se usar o comando `imp`. Por último, o comando `imp$imp$NOME DA VARIÁVEL` mostra todos os valores que foram gerados em todas as ***m*** imputações realizadas.

Para a imputação múltipla pelo método BLR somente o comando `imp` sofre modificação, ficando: `imp <- mice(imputacao, defaultImp = c("norm")).`

ARTIGO 2

COMPARAÇÃO DE MÉTODOS DE IMPUTAÇÃO ÚNICA E MÚLTIPLA APLICADOS A UM MODELO DE RISCO PARA MORTALIDADE CIRÚRGICA

COMPARISON OF SIMPLE AND MULTIPLE IMPUTATION METHODS APPLIED TO A RISK MODEL FOR SURGICAL MORTALITY

Luciana Neves Nunes, ***doutoranda em Epidemiologia pela UFRGS.***

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL (UFRGS)

A ser enviado a Revista Brasileira de Epidemiologia

ARTIGO 2

COMPARAÇÃO DE MÉTODOS DE IMPUTAÇÃO ÚNICA E MÚLTIPLA APLICADOS A UM MODELO DE RISCO PARA MORTALIDADE CIRÚRGICA

COMPARISON OF SIMPLE AND MULTIPLE IMPUTATION METHODS APPLIED TO A RISK MODEL FOR SURGICAL MORTALITY

Luciana Neves Nunes[1,2] , Mariza Machado Klück[1,3] e Jandyra Maria Guimarães Fachel[1,2]

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

2. Departamento de Estatística, Instituto de Matemática, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil.

3. Departamento de Medicina Social, Faculdade de Medicina, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil

Endereço para correspondência

Luciana Neves Nunes (lununes@mat.ufrgs.br)

Universidade Federal do Rio Grande do Sul

Instituto de Matemática - Departamento de Estatística

Av. Bento Gonçalves, 9500

Bairro Agronomia

Porto Alegre – RS

CEP: 91509-900

Resumo

Introdução: Perda de informações é um problema freqüente em estudos que são 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 estudo real 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.

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: A suposição MAR parece adequada, pois os resultados indicam que faz diferença levar em conta a relação da albumina com outras variáveis observadas. A imputação única subestima a variabilidade, gerando intervalos de confiança muito estreitos. A IM se mostrou melhor que a imputação única, pois leva em consideração 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 the imputation complete data sets are artificially constructed and can be analised by traditional statistical analysis. The objective of this paper is to compare three types of imputation.

Methods: The data used come from a study for the development of risk models for surgical mortality. The sample size was 450 patients. The applied imputation methods were: two single imputation and one multiple imputation and the assumption was MAR.

Results: The variable with missing data was the serum albumin with 27,1% of missing. The logistic models adjusted with the imputed data by the simple imputation were similar, but differed from models obtained with imputed data by the multiple imputation in relation to the inclusion of variables.

Conclusions: The MAR assumption seems adequate since different results are obtained if the relation among the albumin and the other observed variables is considered. The single imputation underestimate the variability generating confidence intervals too narrow. The multiple imputation was better than the simple imputatios since the multiple imputation account for the variability among imputations for the model estimates.

Key-words: Imputation methods; multiple imputation; missing data; missing at random.

Introdução

Perda de informações é um problema freqüente 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].

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 que, após a imputação dos dados, o investigador pode utilizar técnicas tradicionais de análise estatística para dados completos[2,3,4].

Métodos simples como imputação pela média ou pela mediana, também conhecidos como métodos de imputação única, têm sido bastante usados pela sua facilidade de implementação. Entretanto existem desvantagens na utilização desses métodos, como a subestimação da variabilidade e a impossibilidade da utilização de outras variáveis do próprio conjunto de dados para melhorar o processo de imputação, ou seja, possíveis relações entre variáveis não são levadas em conta[1,4,5].

Como alternativa à imputação única e com o objetivo de corrigir suas desvantagens, surgiu na década de 80 a imputação múltipla proposta por Donald Rubin[6,7]. A idéia 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[1,2,5].

Embora a imputação múltipla tenha boas propriedades estatísticas, ela ainda não é usada com freqüência na área da saúde[1,3,4]. Nesse 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.

Método

Fonte de dados

O banco de dados utilizado nesse 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[8] 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. Uma descrição das variáveis de interesse para esse trabalho é apresentada na Tabela 1. As variáveis apresentadas fazem parte do modelo final obtido por Klück[8].

Tabela 1 – Descrição das variáveis utilizadas

Variável	Categorias	Descrição/Observação
Óbito	Sim	Óbito num período de até 30 dias após a realização da cirurgia.
	Não	
ASA*	I	Paciente saudável, sem doença sistêmica e fora dos extremos de idade.
	II	Indivíduo com uma doença sistêmica bem controlada, que não afeta sua atividade diária ou paciente com um risco anestésico como tabagismo, obesidade ou alcoolismo.
	III	Indivíduo com múltiplas doenças sistêmicas ou com uma doença sistêmica grave, que limite sua atividade diária.
	IV	Indivíduo com doença severa e incapacitante, em estágio terminal, ou mal controlada.
	V	Paciente em iminente risco de morte, sendo a cirurgia o último recurso possível para preservar a vida ou atenuar o sofrimento.
Idade	< 75 anos	
	≥ 75 anos	
Caráter da cirurgia	Eletiva	
	Urgência	
Albumina	≤ 2,2 g/dl	
	2,3 a 3,0 g/dl	Obs: Variável que apresentou 27,1% de dados faltantes.
	> 3,0 g/dl	

*Avaliação pré-anestésica segundo a *American Society of Anestesiology (ASA)*.

A partir da escolha das variáveis de interesse, foram feitos os três tipos de imputações para a albumina que foi a variável que teve dados faltantes. As duas imputações únicas foram feitas da seguinte maneira: a) 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 b) 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 80, Rubin[6] 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 imputações nos resultados, tornando as estimativas mais eficientes[2,6]. Ilustrativamente, a IM pode ser representada com na figura 1:

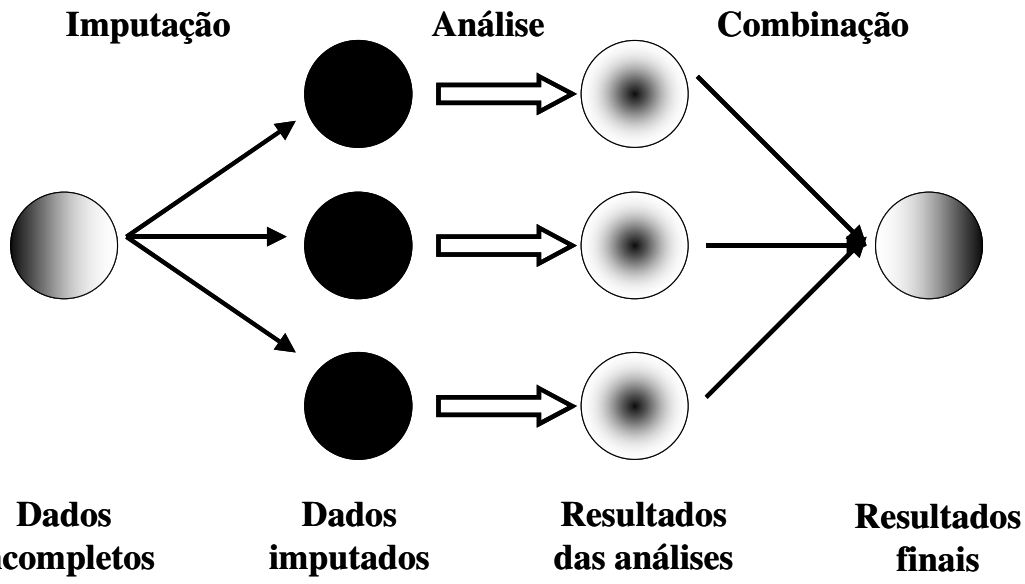

Figura 1 – Esquema da imputação múltipla (Figura extraída de www.multiple-imputation.com)

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[2,6,9,10,11].

Nesse trabalho far-se-á a suposição de que os dados faltantes da albumina seguem o mecanismo de perdas ao acaso MAR, ou seja, de que os dados faltantes podem ser previstos a partir de outras variáveis observadas e padrão monotônico[11].

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 proposto por Rubin[6].

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.[2,6]

As regras de Rubin podem ser descritas como 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 ou razão de chances[5,12]. 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 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: $T = \overline{U} + \left(1 + \frac{1}{m}\right)B$ [2,6].

A imputação múltipla neste trabalho foi feita pela chamada *Bayesian Linear Regression* (*BLR*) (Pág.166-167, Rubin, 1987)[6]. Foi escolhido este método de imputação múltipla por ser adequado para 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[14] que foi usado para esse trabalho. Mais detalhes computacionais tais como as instruções de programação podem ser encontradas em Nunes[15] ou no manual do pacote.[12]

Como o método de IM usado neste trabalho é baseado numa análise de regressão linear, foi necessário definir as variáveis que seriam usadas para a imputação. Sob a suposição de que os dados faltantes têm mecanismo de não-resposta MAR, as variáveis incluídas no modelo de imputação foram ASA e caráter da cirurgia por apresentarem relação com a variável albumina. A variável resposta é a albumina (Y_{imp}), ou seja, variável a ser imputada. Foram feitas duas imputações múltiplas diferentes, uma não incluindo a variável desfecho do estudo original sobre escore de risco (óbito), e outra sim, chamadas de IM(1) e IM(2), respectivamente. Estas imputações podem ser descritas da seguinte maneira:

IM(1): Albumina = β_1(ASA III)+β_2(ASA IV/V) + β_3(Cirurgia urgente) + Constante

IM(2): Albumina = β_1(ASA III)+β_2(ASA IV/V) + β_3(Cirurgia urgente) + β_4(Óbito=sim) + 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 conforme Tabela 1, sendo a categoria (≥3,1g/dl) a de referência. Depois de realizadas as imputações múltiplas, as estimativas gerais para os coeficientes β foram obtidas pela aplicação das Regras de Rubin citadas anteriormente. As comparações foram realizadas pela comparação dos valores das estimativas pontuais das razões de chance (RC), respectivos erros padrão e intervalos de confiança.

Para realizar as imputações múltiplas foi utilizado o pacote MICE, do aplicativo R[14] versão 2.5.1, sendo que foram feitas m=5 imputações. Para o ajuste das regressões logísticas foi usado o SPSS 13.0[16] e para os cálculos das Regras de Rubin foi utilizado o Excel[17].

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.

Tabela 2 – Comparação entre os coeficientes do modelo de Regressão Logística obtidos com diferentes imputações dos valores faltantes da Albumina.

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

*IM(1): Albumina = β_1(ASA III)+β_2(ASA IV/V) + β_3(Cirurgia urgente) + Constante
**IM(2): Albumina = β_1(ASA III)+β_2(ASA IV/V) + β_3(Cirurgia urgente) + β_4(Óbito=sim)+Constante

Na Tabela 3 podem ser vistos os resultados das estimativas das razões de chance, dos respectivos intervalos de 95% de confiança e erros padrão 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 pela tabela 3 pode se ver que os erros padrão para as categorias da variável albumina foram maiores nas estratégias das imputações múltiplas.

Tabela 3 – Estimativas da RC da regressão logística após as imputações.

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]	3,3[1,4;7,8]	3,6[1,5;8,4]	3,6[1,5;8,5]
	(0,438)	(0,438)	(0,439)	(0,441)
ASA IV/V	22,3[9,5;52,6]	23,0[9,8;54,0]	25,3[10,6;60,4]	24,8[10,3;59,7]
	(0,438)	(0,435)	(0,444)	(0,448)
Idade ≥ 75 anos	4,1[1,8;9,1]	4,1[1,8;9,1]	4,1[1,9;9,2]	4,2[1,9;9,3]
	(0,409)	(0,407)	(0,407)	(0,412)
Alb até 2,2 g/dl	6,6[2,9;14,8]	5,8[2,8;11,8]	4,5[2,0;10,0]	5,0[2,1;11,8]
	(0,414)	(0,363)	(0,407)	(0,437)
Alb 2,3 a 3 g/dl	2,2[1,0;4,7]	2,2[1,1;4,6]	1,7[0,7;3,8]	1,7[0,8;3,8]
	(0,389)	(0,357)	(0,417)	(0,403)

*IM(1): Albumina = β_1(ASA III)+β_2(ASA IV/V) + β_3(Cirurgia urgente) + Constante
**IM(2): Albumina = β_1(ASA III)+β_2(ASA IV/V) + β_3(Cirurgia urgente) + β_4(Óbito=sim)+Constante

Com a idéia 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 nas figuras 2 a 4 e tabelas 4 a 6.

A Figura 2 mostra como ficou a distribuição da variável albumina com as três categorias citadas na Tabela 1, 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 diferentes abordagens para a imputação múltipla IM(1) e IM(2).

Pela figura 2 é possível se observar que a categoria “até 2,2 g/dl” teve menor freqüência na situação da imputação pelas medianas, quando comparada com o grupo sem imputação, o mesmo ocorrendo para a imputação única do valor normal. Ainda, quando observada a categoria “2,3 a 3,0 g/dl”, a imputação pelo valor normal também teve menor freqüência 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 1 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 freqüência 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 foram bastante próximas da situação “sem imputação". Salientando que se tem a suposição de que os dados faltantes ocorreram ao acaso (MAR), 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.

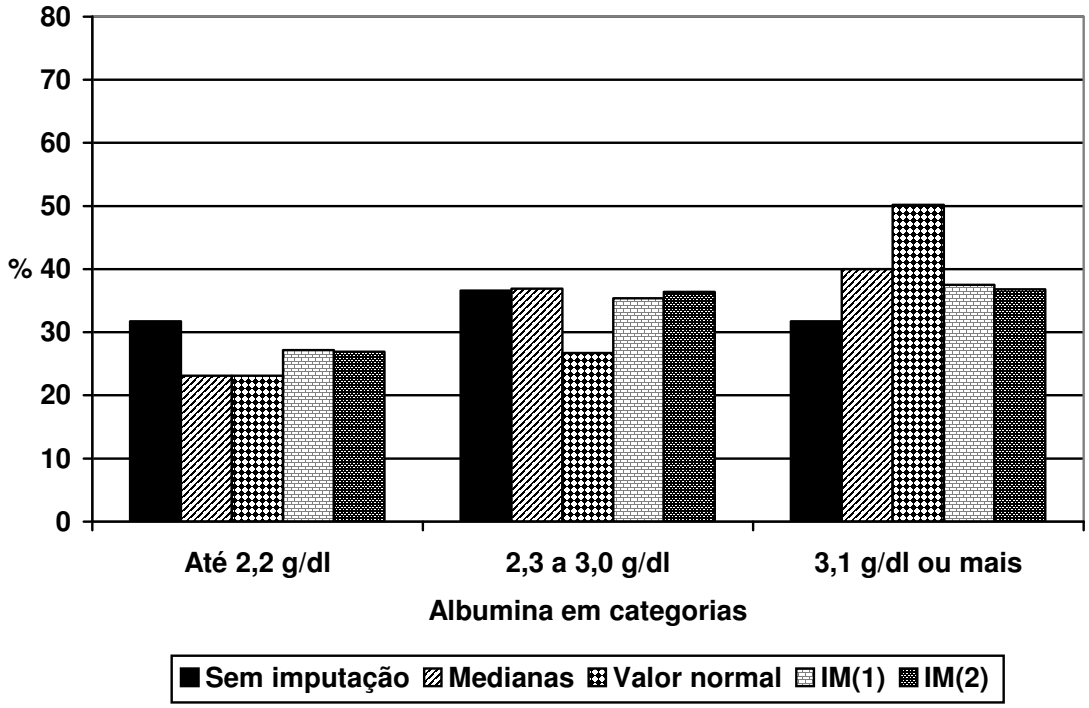

Figura 2 – Comparação das freqüências relativas das categorias de albumina, considerando os dados sem imputação, imputados pelas medianas, imputados pelo limite inferior do valor normal (3,5 g/dl) e imputados pelas imputações múltiplas IM(1) e IM(2).

Para um maior detalhamento dos resultados, as figuras 3 e 4 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 3) e pacientes com cirurgia de urgência (figura 4). Enquanto as IM tiveram distribuições bastante semelhantes à distribuição com dados faltantes (sem imputação), tanto na figura 3 como na 4 as imputações únicas apresentaram discrepâncias maiores. Por exemplo, na figura 3, as freqüências 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 4) as maiores discrepâncias foram a freqüência da imputação pela mediana na categoria "2,3 a 3,0

g/dl" e a freqüência da imputação do valor normal na categoria "3,1 g/dl ou mais", ambas maiores em relação aos dados sem imputação.

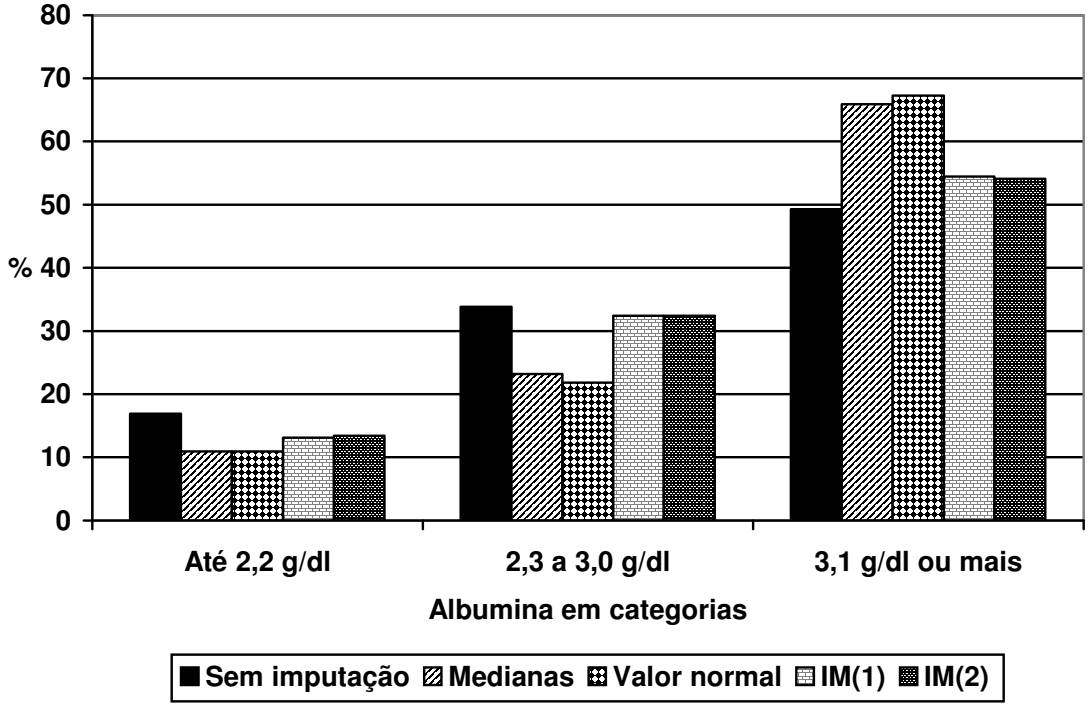

Figura 3 – Comparação das freqüências relativas das categorias de albumina, considerando os dados sem imputação, imputados pelas medianas, imputados pelo limite inferior do valor normal (3,5 g/dl) e imputados pelas imputações múltiplas IM(1) e IM(2) para o grupo de Cirurgia Eletiva.

As médias e desvios padrões dos valores da albumina com dados faltantes e com as diferentes imputações usadas no trabalho são apresentados na tabela 4. Todas as abordagens mostraram resultados bastante semelhantes, mas as imputações únicas tiveram os desvios padrões menores.

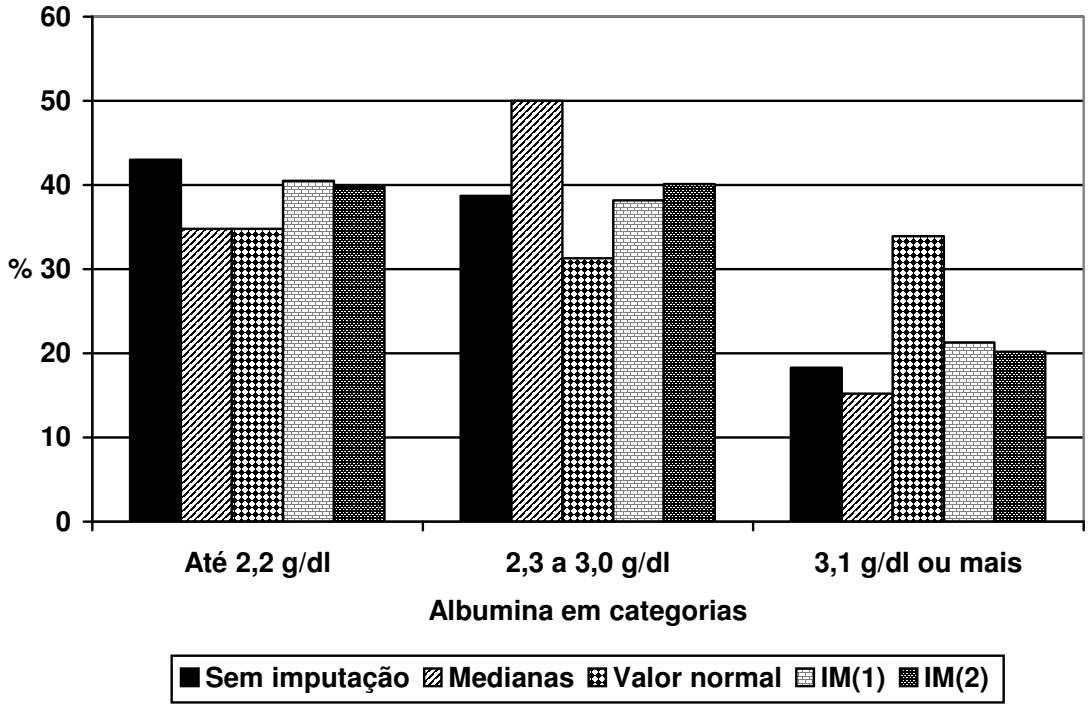

Figura 4 – Comparação das freqüências relativas das categorias de albumina, considerando os dados sem imputação, imputados pelas medianas, imputados pelo limite inferior do valor normal (3,5 g/dl) e imputados pelas imputações múltiplas IM(1) e IM(2) para o grupo de Cirurgia Urgente.

Tabela 4 – Médias e desvios padrões da variável albumina nas diferentes formas de imputações de dados

Método de imputação	Média	DP
Sem imputação	2,702	0,781
Medianas	2,738	0,693
Valor normal (3,5 g/dl)	2,918	0,756
IM(1)	2,803	0,808
IM(2)	2,780	0,788

A tabela 5 mostra como os dados faltantes da albumina estão distribuídos em relação as variáveis ASA e caráter de cirurgia. No banco de dados, 122 pacientes não tinham a informação da albumina, ou seja, 27,1% dos pacientes. 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%. Observando-se os resultados dos cruzamentos, percebe-se que do total de faltantes, 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 tiveram cirurgia de urgência que pertenciam à ASA IV/V. Além disso, há uma associação positiva entre valor baixo de ASA (I/II) e cirurgia eletiva (χ^2=17,64; gl=2; p<0,001). Ou seja, a maior falta de dados de albumina nesta categoria não é independente do tipo de cirurgia: há maior número de dados de albumina faltantes se a cirurgia é eletiva.

Tabela 5 – Padrão dos dados faltantes de albumina (n=122) em relação a ASA e caráter de cirurgia. Percentual entre parênteses.

	Caráter de cirurgia		
ASA	Eletiva n (%)	Urgência n (%)	Total 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)

(χ^2=17,64; gl=2; p<0,001)

Tabela 6 – Medianas da variável albumina (em g/dl), conforme ASA e caráter de cirurgia (n=328)

	Caráter de cirurgia		
ASA	Eletiva	Urgência	Total
ASA I/II	3,2	2,7	3,0
ASA III	3,1	2,6	2,8
ASA IV/V	2,2	2,1	2,1
Total	3,1	2,4	2,6

A tabela 6 mostra os valores das medianas de acordo com o cruzamento das categorias de ASA e caráter de cirurgia. Percebe-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 as 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

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[18].

Segundo Ambler et al.[18], muitos modelos de risco da área médica usam dados rotineiramente coletados em hospitais. Tais conjuntos de dados contêm informações dos preditores baseadas nas características dos pacientes. Desfechos clínicos como morte hospitalar, por exemplo, em geral são informações completas, entretanto muitos dos preditores têm observações faltantes. É freqüente que pacientes tenham vários preditores sem informações e, ainda, não é incomum que alguns preditores importantes tenham mais do que 50% de dados faltantes[19]. 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[20].

Usualmente, modelos de risco são estimados através de análise de regressão logística. Muitos pesquisadores restringem a análise aos casos completos ignorando os dados faltantes. Essa abordagem exclui todos os pacientes que tenham a informação

incompleta em qualquer um dos preditores. Também pode acontecer que um preditor seja excluído do processo de modelagem caso tenha muitos pacientes sem a informação. Ainda, analisar somente os casos completos pode resultar em tamanhos de amostras menores que o planejado. Com esses problemas, os modelos ajustados podem ter um pior poder preditivo. 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[21].

Apesar de, atualmente, a estratégia de imputação de dados já estar bastante difundida, a utilização da imputação múltipla ainda é muito insipiente, principalmente na área da saúde.[1] Talvez isso aconteça pela complexidade computacional que a imputação múltipla exige, principalmente quando comparada com os métodos de imputação única, tais como os apresentados nesse trabalho. Outra razão para sua pouca utilização é que a IM pode ser bastante complexa dependendo de suposições quanto às possíveis justificativas para o surgimento de não-respostas nos conjuntos de dados.[2]

Neste trabalho foi feita uma avaliação de alguns métodos de imputação que são usados no contexto de desenvolvimento de modelos de risco, através da investigação do desempenho de três métodos de imputação, sendo dois de imputação única e um de IM. 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 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 dos dados analisados, isso causaria sérias implicações clínicas[18].

Os métodos de imputação única, que são regularmente usados na prática provavelmente pela sua simplicidade, normalmente mostram um ganho em relação à

análise restrita aos casos completos. Entretanto, esses métodos podem reduzir a variabilidade amostral por imputarem valores do centro da distribuição (método das medianas) ou imputarem um único valor (método do valor normal) para todos os pacientes com dados faltantes[22].

Uma vantagem da IM em relação à imputação única é que a IM leva em conta a variabilidade entre imputações, e por ter o componente Bayesiano embutido no procedimento restringe 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[18].

Como a imputação única não permite levar em conta a variabilidade entre imputações, os IC´s dos coeficientes de regressão podem ser muito estreitos. Já a IM tem a vantagem de levar em conta a incerteza das imputações[22].

Segundo F. E. Harrel Jr.[21], quando há mais que 15% de perda de dados, a imputação múltipla é indicada na maior parte dos modelos. Portanto, como nesse trabalho havia 27,1% de dados faltantes para a variável albumina, utilizou-se esse critério para justificar a aplicação da IM.

A idéia usada neste artigo foi a de comparar o método de imputação múltipla *BLR* (*Bayesian Linear Regression)* com dois métodos de imputação única através de um estudo real em que se ajustou um modelo de risco para mortalidade cirúrgica. Esta comparação foi feita de maneira simples, usando como medidas de comparação as estimativas de um modelo de regressão logística após a realização das diferentes imputações. Os modelos ajustados com os bancos de dados completados pelas imputações incluíram sempre as mesmas variáveis com as mesmas categorizações para que fosse possível a comparação dos resultados obtidos.

Um aspecto importante da IM é a inclusão das variáveis nos modelos de imputações, ou seja, a IM pode facilmente levar em conta mais informações que as

imputações únicas[18]. Neste trabalho, a escolha pelo método BLR foi ditada pela simplicidade de seu uso e por ser adequado para os dados do exemplo.

Na técnica *BLR*, é preciso que se faça a escolha de variáveis que serão consideradas para se fazer a imputação. De acordo com Van Buuren et al.[13,23] , pode-se adotar como regra geral usar toda a informação das variáveis disponíveis, o que produziria imputações múltiplas com mínimo viés e máxima precisão. Alguns autores observam que incluir tantos preditores quanto possível tende a fazer a suposição de MAR mais plausível[13,23].

Entretanto, é preciso ser parcimonioso na inclusão dessas variáveis para a IM para se evitar problemas computacionais e de multicolinearidade. Não é necessário que se incluam todas as variáveis que estão disponíveis num banco de dados para a imputação, mas apenas aquelas que potencialmente podem ter boa capacidade preditiva.[16] Em geral, o modelo de imputação deve conter variáveis com potencial poder preditivo para o valor perdido e acomodar a estrutura, por exemplo, de interações. Falha em acomodar a estrutura pode gerar viés nos resultados[3].

A suposição MAR ocupa uma importante posição na questão de tratamento de dados faltantes, 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[3]. Infelizmente, é impossível determinar se os dados faltantes são MAR ou NMAR, isso pode ser simplesmente especulado[4].

Quando os dados são NMAR é necessário que se incorpore explicitamente o mecanismo de não-resposta, algo que na maioria das situações se desconhece. Embora a suposição MAR leve a considerável simplificação quando se analisam dados faltantes, infelizmente é uma suposição que não pode ser testada, como se faz com a suposição de normalidade, por exemplo[3]. Assim, é uma característica da análise de dados

incompletos que dependa de suposições que não podem ser testadas. Nesse sentido, tais análises pertencem a uma classe na qual se incluem confundidores não observados, erros de medida e não-adesão ao tratamento.

A partir das idéias expostas sobre a inclusão de variáveis no modelo de imputação e a suposição MAR, incluíram-se as variáveis ASA e caráter da cirurgia nos modelos das IM, considerando-se a relação que apareceu entre essas variáveis e a variável albumina. Por outro lado, seguindo a conclusão de Moons et al.[24] que verificaram ser desejável a inclusão do desfecho no modelo de imputação, foi considerado 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[24].

Considerando-se que a IM forneceu valores plausíveis para os dados faltantes, é interessante comentar mais detalhadamente a diferença observada entre as estimativas dos coeficientes da variável albumina. Os valores obtidos pelas IM podem ser considerados plausíveis pela suposição que se fez sobre o padrão de não-resposta, nesse caso, MAR, ou seja, o padrão de não-resposta pode ser previsto por outras variáveis disponíveis no banco de dados. Nas tabelas 5 e 6 é possível observar-se que o padrão dos dados faltantes da albumina ocorre da seguinte maneira: quanto mais grave a situação do paciente, menos dados faltantes ocorrem e mais baixo o valor da albumina, sugerindo uma relação entre a albumina e as variáveis ASA e caráter da cirurgia.

A imputação múltipla adotada nesse trabalho procurou contemplar a relação existente entre albumina, categorias de ASA e caráter da cirurgia, pois essas foram as variáveis incluídas no método BLR usado para se fazer as IM. Como as imputações únicas que foram apresentadas não contemplam essa relação, isso provavelmente fez com que as estimativas dos modelos fossem diferentes.

O método das medianas levou em conta o caráter da cirurgia, entretanto 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 somente 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 figuras 2 a 4).

Essas diferenças nas imputações única e múltipla podem ter levado aos diferentes modelos obtidos, no sentido de que a categoria "2,3 a 3,0 g/dl" da albumina, enquanto é significativa nos modelos com os dados imputados pelas imputações únicas, deixa de ser 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 a dados faltantes[3].

Também é interessante ressaltar que quanto à variabilidade das estimativas, os resultados foram de acordo com o que se encontra na literatura[5,7], ou seja, de que imputações únicas subestimam a variabilidade, gerando erros padrão menores que as imputações múltiplas.

Esse trabalho mostrou a importância de se utilizar a imputação múltipla, que permite considerar o padrão de não-resposta MAR e a relação entre a variável com dados faltantes e outras variáveis observadas. Cabe ressaltar que, se esta relação não for preservada, a inferência pode ficar viesada[2,5]. No entanto, deve haver cuidado na generalização dos resultados obtidos com esse trabalho, já que eles foram obtidos em uma situação particular, quanto a tamanho amostral, tipo de variável e estrutura de

relação entre as variáveis envolvidas. Mais trabalhos são necessários para avaliar o desempenho de métodos de imputação para conjuntos de dados com maior proporção de preditores contínuos com possível relação não-linear e associação mais forte entre os preditores, bem como conjuntos de dados com desfechos não-binários[18].

Referências bibliográficas

1. Zhou, XH, Eckert, GJ and Tierney, WM. Multiple imputation in public health research. **Statistics in Medicine**, 2001; 20(9-10):1541-49.
2. Tang L, Song J, Belin TR, Unützer J. A comparison of imputation methods in a longitudinal randomized clinical trial. **Statistics in Medicine**, 2005; 24: 2111-28.
3. Kenward MG e Carpenter J. Multiple imputation: current perspectives. **Statistical Methods in Medical Research**, 2007; 16(3): 199-218.
4. van der 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. **Journal of Clinical Epidemiology**, 2006; 59(10): 1102-1109.
5. Schafer JL. Multiple imputation: a primer. **Statistical Methods in Medical Research** 1999; 8(1): 3-15.
6. Rubin DB. **Multiple Imputation for Nonresponse in Surveys.** New York: Wiley; 1987.
7. Rubin DB. Multiple imputation after 18+ years. **Journal of the American Statistical Association** 1996; 91: 473-89.
8. Klück M **Metodologia para ajuste de indicadores de desfechos hospitalares por risco prévio do paciente**. Tese de doutorado em Medicina: Epidemiologia.

Faculdade de Medicina. Universidade Federal do Rio Grande do Sul , Porto Alegre 2004.

9. Little RJA. Regression with Missing Xs - A Review. **Journal of the American Statistical Association** 1992; 87(420): 227-37.
10. Zhang P. Multiple imputation: Theory and method. **International Statistical Review** 2003; 71(3): 581-92.
11. Schafer, JL and Graham JW. Missing data: our view of the state of the art. **Psychological Methods** 2002; 7: 147-177.
12. Bernaards AB, Farmer MM, Qi K, Dulai GS, Ganz PA and Kahn KL. Comparison of two multiple imputation procedures in a cancer screening survey. **Journal of Data Science**, 2003; 1: 293-312.
13. Van Buuren S and Oudshoorn CGM. **Multivariate imputation by chained equations. MICE V1.0 User's Manual**. Report PG/VGZ/00.038. Leiden: TNO Preventie en Gezondheid, 2000.
14. R Development Core Team R. A language and environment for statistical computing. Vienna, Austria: R **Foundation for Statistical Computing**; 2004. Available at http://www.R-project.org ISBN 3-900051-00-3, Accessed September 02, 2007.
15. Nunes, LN. **Métodos de imputação de dados aplicados na área da saúde**. Tese de doutorado em Medicina: Epidemiologia. Faculdade de Medicina. Universidade Federal do Rio Grande do Sul , Porto Alegre 2007.
16. SPSS for Windows. Release 13.0.1, 2004. SPSS Inc. 1989-2004.
17. Microsoft® Excel 2002. Copyright© Microsoft Corporation1985-2001.

18. Ambler G, Omar RZ, Royston P. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. **Statistical Methods in Medical Research**, 2007; 16(3): 277-298.

19. 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.

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

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

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

23. Van Buuren S, Boshuizen HC and Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. **Statistics in Medicine** 1999; 18:681-694.

24. Moons KG, Donders RA, Stijnen T, Harrell FE Jr. Using the outcome for imputation of missing predictor values was preferred. **Journal of Clinical Epidemiology**, 2006; 59(10): 1092-1101.

6. CONCLUSÕES E CONSIDERAÇÕES FINAIS

O objetivo dessa tese foi divulgar a metodologia de imputação de dados, incluindo a imputação múltipla. Foi gerado um guia para a aplicação da metodologia para usuários não estatísticos da área da saúde, foi feita uma aplicação do método de imputação múltipla a conjuntos de dados reais e foram feitas comparações entre os resultados obtidos com e sem os dados imputados.

Vários trabalhos foram publicados sobre métodos de imputação múltipla, inclusive na área da saúde. No entanto, muito pouco sobre esse assunto tem aparecido na literatura brasileira, tornando interessante o presente trabalho.

Nas duas últimas décadas tem havido intensa produção científica internacional sobre a metodologia de imputação múltipla. Cada vez mais têm sido publicados artigos sobre o problema da não-resposta e propostas de técnicas para solucioná-lo. Inclui-se aí a imputação múltipla, que tem aparecido tanto como simples aplicação como desenvolvimento de novas metodologias.

Através dos dois artigos apresentados nesse trabalho mostrou-se que é computacionalmente viável que se faça o uso da imputação múltipla.

Outro ponto relevante desse trabalho foi que através da aplicação de diferentes métodos de imputação a um banco de dados reais, foi possível mostrar que a imputação múltipla tem vantagens sobre a imputação única.

Outro aspecto positivo desse trabalho foi a verificação de que o pacote R tem a facilidade de ter um programa desenvolvido especificamente com o fim de se fazer imputações múltiplas, o MICE. O programa mostrou-se relativamente simples de se usar, exigindo apenas um conhecimento básico do pacote R para se fazer as análises

pretendidas. O MICE permite ainda realizar outros tipos de imputações múltiplas e as sintaxes para a implementação podem ser obtidas no manual.

Fica como sugestão para trabalhos futuros a apresentação de outros métodos de imputação múltipla, por exemplo, aqueles específicos para variáveis não-numéricas.

7. ANEXOS

PROJETO

Questão/Objetivo Geral

Este trabalho tem por objetivo descrever e aplicar os métodos estatísticos utilizados para Imputação de Dados, ou seja, substituição dos dados faltantes em bancos de dados epidemiológicos e em geral. Daremos especial ênfase ao método de imputação denominado Imputação Múltipla, que utiliza recursos computacionais modernos como Bootstrap e MCMC (Markov Chain Monte Carlo). O primeiro artigo será de caráter metodológico, e incluirá a descrição dos diversos métodos de imputação de dados e a comparação destes quando for conveniente e/ou necessário. Será feita ampla revisão de literatura e serão apresentados detalhes técnicos e computacionais para que o leitor possa reproduzir a técnica em seus dados. O segundo artigo mostrará a aplicação da metodologia em um banco de dados reais, extraído dos Protocolos do HCPA e já utilizado em tese de doutorado recente (Kluck, 2004).

Justificativa

Uma complicação comum em investigações científicas é a ocorrência de dados ausentes (missing data), especialmente na área da Saúde e das Ciências Sociais (Rubin, 1996). Determinar a abordagem analítica adequada para bancos 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 conjunto de dados. O desenvolvimento de métodos estatísticos direcionados a solucionar problemas de dados ausentes tem sido uma área de pesquisa bastante ativa nas últimas décadas (Rubin, 1976; Little and Rubin, 1987; Little, 1992; 1995; Schafer, 1997, 1999; Zhang, 2003).

Em situações com dados ausentes, uma abordagem bastante comum é restringir a análise aos sujeitos com dados completos nas variáveis envolvidas. Porém, as estimativas obtidas com tais análises podem ser viesadas, especialmente se os indivíduos que são incluídos na análise são sistematicamente diferentes daqueles que foram excluídos em termos de uma ou mais variáveis. Para contornar esse problema, desde os anos 80 surgiram técnicas estatísticas que envolvem imputação de dados

ausentes. Essas técnicas têm por objetivo “completar” os bancos de dados 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 ausentes pela média, pela mediana, por interpolação ou até por regressão linear. Todas essas técnicas mencionadas permitem “preencher” os dados ausentes através 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 o banco completo sejam válidos, pois os valores imputados não são valores reais.

Rubin, ainda nos anos 70, propôs uma técnica chamada *imputação múltipla* (*IM*) para resolver o problema de não-resposta em pesquisas. No entanto, apenas recentemente esta técnica vem sendo mais utilizada devido aos desenvolvimentos computacionais para implementação da técnica. Essa técnica possibilita a inclusão da incerteza da imputação nos resultados, corrigindo o maior problema associado à imputação única. A *IM* consiste de três passos:

1) São obtidos ***m*** bancos de dados completos através 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 dois são combinados de um jeito simples e apropriado para se obter a chamada inferência da imputação repetida.

O primeiro passo é a parte fundamental da *IM*, pois as técnicas de imputação utilizadas têm que preservar a relação das observações ausentes e presentes e ainda levar em conta o mecanismo de ausência e o padrão dos dados ausentes. Os mecanismos 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*); e os padrões são: monótono e não-monótono.

Considerados o mecanismo e o padrão os métodos de imputação múltipla são:

1) Escore da tendência *(Propensity Score Method*)
2) Modelo preditivo (*Predictive Model Method*)
3) MCMC (*Markov Chain Monte Carlo*)

Para a realização da análise computacional alguns pacotes têm sido bastante referidos na literatura, pois suportam muito bem o uso dos métodos citados. Dentre os

mais utilizados pode-se citar o SAS, S-Plus, SOLAS, NORM, BMDP e MICE, sendo que o MICE é de domínio público, fato que facilita sua utilização. Uma análise do desempenho dos pacotes computacionais para *IM* pode ser vista em Horton e Lipsitz (2001).

Devido ao crescimento do interesse por essa área, muitas publicações têm sido feitas nos últimos anos. Uma busca no PubMed com a palavra chave “imputation” indicou 238 trabalhos publicados (06/10/2005) no período de 2000 a 2005. Por exemplo, a revista *Statistics in Medicine*, importante periódico tanto na Medicina como na Estatística, dedicou os fascículos 1 a 3 do volume 16 para tratar de dados ausentes. Estes fatos revelam que o estudo de metodologias para dados ausentes está bastante debatido atualmente, o que indica a relevância desse trabalho.

REFERÊNCIAS BIBLIOGRÁFICAS:

Engels, J. M. and P. Diehr. "Imputation of missing longitudinal data: a comparison of methods." Journal of Clinical Epidemiology 56.10 (2003): 968-76.

Horton, N. J. and S. R. Lipsitz. "Multiple imputation in practice: Comparison of software packages for regression models with missing variables." American Statistician 55.3 (2001): 244-54.

Little, R. J. A. "Regression with Missing Xs - A Review." Journal of the American Statistical Association 87.420 (1992): 1227-37.

Nielsen, S. F. "Proper and improper multiple imputation." International Statistical Review 71.3 (2003): 593-607.

Rubin, D. B. "Multiple imputation after 18+ years." Journal of the American Statistical Association 91.434 (1996): 473-89.

Schafer, J. L. "Multiple imputation: a primer." Statistical Methods in Medical Research 8.1 (1999): 3-15.

van Buuren, S., H. C. Boshuizen, and D. L. Knook. "Multiple imputation of missing blood pressure covariates in survival analysis." Statistics in Medicine 18.6 (1999): 681-94.

Zhang, P. "Multiple imputation: Theory and method." International Statistical Review 71.3 (2003): 581-92.

Zhou, X. H., G. J. Eckert, and W. M. Tierney. "Multiple imputation in public health research." Statistics in Medicine 20.9-10 (2001): 1541-49.

LINHA DE PESQUISA PRÉVIA

Métodos Estatísticos Aplicados em Epidemiologia

Planejamento da Pesquisa

Essa parte do anteprojeto não se aplica, pois os dados que serão utilizados para a análise serão secundários.

Questões Éticas

As questões éticas serão avaliadas pelo comitê de ética da do Hospital de Clínicas de Porto Alegre, pois os dados que serão utilizados serão de pacientes de lá. Os dados serão secundários e não haverá forma de identificar os pacientes. Devido ao fato dos dados serem secundários e anônimos, neste trabalho não há questões éticas importantes.

Cronograma Básico

6 meses – Revisão de literatura

8 meses – Parte computacional: simulação e análise de dados reais

6 meses – Redação final

Recursos Necessários

- Banco de dados
- Computador para simulação e análise dos dados reais.

Problemas/Soluções/Tarefas

Problema: Obtenção do banco de dados.

Solução: Contatar responsável por dados no Hospital de Clínicas.