Universidade do Estado do Rio de Janeiro

Centro Biomédico

Instituto de Medicina Social

Thaís de Paulo Rangel

Imputação múltipla de dados faltantes:

Exemplo de aplicação no Estudo Pró-Saúde

Rio de Janeiro

2013

Thaís de Paulo Rangel

Imputação múltipla de dados faltantes:
Exemplo de aplicação no Estudo Pró-Saúde.

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

Orientadores: Prof. Dr. Eduardo Faerstein

Prof. Dr. Washington Leite Junger

Rio de Janeiro

2013

CATALOGAÇÃO NA FONTE
UERJ/REDE SIRIUS/CB/C

R196 Rangel, Thaís de Paulo
Imputação múltipla de dados faltantes : exemplo de aplicação no Estudo Pró-Saúde / Thaís de Paulo Rangel. – 2013.
150f.

Orientador: Eduardo Faerstein.
Coorientador: Washington Leite Junger

Dissertação (mestrado) – Universidade do Estado do Rio de Janeiro, Instituto de Medicina Social.

1. Epidemiologia –Modelos estatísticos - Teses. 2. Epidemiologia - Metodologia – Teses. 3. Epidemiologia - Pesquisa – Teses. I. Faerstein, Eduardo. II. Junger, Washington Leite. III. Universidade do Estado do Rio de Janeiro. Instituto de Medicina Social. III. Título.

CDU 616-036.22:001.8

_____ _____

Assinatura Data

Thaís de Paulo Rangel

Imputação múltipla de dados faltantes:

Exemplo de aplicação no Estudo Pró-Saúde

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

Aprovada em 05 de março de 2013.

Orientadores: Prof. Dr. Eduardo Faerstein
Instituto de Medicina Social – UERJ

Prof. Dr. Washington Leite Junger
Instituto de Medicina Social – UERJ

Banca Examinadora: _____

Prof. Dr. Cláudio José Struchiner
Instituto de Medicina Social – UERJ

_____

Prof.ª Dra. Luciana Neves Nunes
Universidade Federal do Rio Grande do Sul - UFRGS

_____

Prof. Dr. Oswaldo Gonçalves Cruz
Programa de Computação Científica – PROCC/Fiocruz

Rio de Janeiro
2013

DEDICATÓRIA

Aos meus pais.

Ao meu avô, Benedicto F. Rangel (*in memorian*).

AGRADECIMENTOS

Aos meus pais, pelo apoio incondicional.

Ao Bernardo, por me encorajar nas horas mais difíceis e me acompanhar em todos os momentos.

Aos meus orientadores Eduardo Faerstein e Washington Junger, por terem confiado em mim desde o primeiro momento e por auxiliarem pacientemente ao longo de todo o desenvolvimento da dissertação.

Aos professores Cláudio Struchiner, Luciana Nunes e Oswaldo Cruz, pela disponibilidade em participar da banca da defesa desta dissertação.

A todos os professores do IMS que contribuíram com meu aprimoramento em Epidemiologia.

Às amigas do mestrado, que desde a primeira aula ajudaram a passar pelos momentos mais complicados nesta caminhada – especialmente Marcela Ferreira e Débora França, por oferecerem um ombro amigo sempre que necessário.

Aos familiares e amigos mais próximos, por compreenderem meu afastamento.

À toda a equipe do Pró-Saúde, especialmente Karine – que me recebeu com muito carinho desde a minha primeira visita ao IMS; e Jaqueline – por todos os conselhos no último ano.

Aos funcionários da secretaria e do laboratório de informática do IMS, por toda a gentileza e paciência sempre que os requisitei.

Aos funcionários participantes do Estudo Pró-Saúde.

À CAPES, pelo financiamento do mestrado.

We cannot stop users from doing bad science, but if possible we should facilitate their ability to do good science with their available tools, even when data sets suffer from missing values.

Donald Rubin

RESUMO

RANGEL, Thaís de Paulo. *Imputação múltipla de dados faltantes*: exemplo de aplicação no Estudo Pró-Saúde. 2013. 150f. Dissertação (Mestrado em Saúde Coletiva) – Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2013.

Dados faltantes são um problema comum em estudos epidemiológicos e, dependendo da forma como ocorrem, as estimativas dos parâmetros de interesse podem estar enviesadas. A literatura aponta algumas técnicas para se lidar com a questão, e, a imputação múltipla vem recebendo destaque nos últimos anos. Esta dissertação apresenta os resultados da utilização da imputação múltipla de dados no contexto do Estudo Pró-Saúde, um estudo longitudinal entre funcionários técnico-administrativos de uma universidade no Rio de Janeiro. No primeiro estudo, após simulação da ocorrência de dados faltantes, imputou-se a variável cor/raça das participantes, e aplicou-se um modelo de análise de sobrevivência previamente estabelecido, tendo como desfecho a história auto-relatada de miomas uterinos. Houve replicação do procedimento (100 vezes) para se determinar a distribuição dos coeficientes e erros-padrão das estimativas da variável de interesse. Apesar da natureza transversal dos dados aqui utilizados (informações da linha de base do Estudo Pró-Saúde, coletadas em 1999 e 2001), buscou-se resgatar a história do seguimento das participantes por meio de seus relatos, criando uma situação na qual a utilização do modelo de riscos proporcionais de Cox era possível. Nos cenários avaliados, a imputação demonstrou resultados satisfatórios, inclusive quando da avaliação de *performance* realizada. A técnica demonstrou um bom desempenho quando o mecanismo de ocorrência dos dados faltantes era do tipo MAR (*Missing At Random*) e o percentual de não-resposta era de 10%. Ao se imputar os dados e combinar as estimativas obtidas nos 10 bancos (m=10) gerados, o viés das estimativas era de 0,0011 para a categoria preta e 0,0015 para pardas, corroborando a eficiência da imputação neste cenário. Demais configurações também apresentaram resultados semelhantes. No segundo artigo, desenvolve-se um tutorial para aplicação da imputação múltipla em estudos epidemiológicos, que deverá facilitar a utilização da técnica por pesquisadores brasileiros ainda não familiarizados com o procedimento. São apresentados os passos básicos e decisões necessárias para se imputar um banco de dados, e um dos cenários utilizados no primeiro estudo é apresentado como exemplo de aplicação da técnica. Todas as análises foram conduzidas no programa estatístico R, versão 2.15 e os scripts utilizados são apresentados ao final do texto.

Palavras-chave: Dados faltantes. Imputação múltipla. Análise de sobrevivência. Tutorial

ABSTRACT

RANGEL, Thaís de Paulo. *Multiple imputation of missing data:* application in the Pro-Saude Study. 2013. 128f. Dissertação (Mestrado em Saúde Coletiva) - Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2013.

Missing data are a common problem in epidemiologic studies and depending on the way they occur, the resulting estimates may be biased. Literature shows several techniques to deal with this subject and multiple imputation has been receiving attention in the recent years. This dissertation presents the results of applying multiple imputation of missing data in the context of the Pro-Saude Study, a longitudinal study among civil servants at a university in Rio de Janeiro, Brazil. In the first paper, after simulation of missing data, the variable color/race of the female servants was imputed and analyzed through a previously established survival model, which had the self-reported history of uterine leiomyoma as the outcome. The process has been replicated a hundred times in order to determine the distribution of the coefficient and standard errors of the variable being imputed. Although the data presented were cross-sectionally collected (baseline data of the Pro-Saude Study, gathered in 1999 and 2001), the following of the servants were determined using self-reported information. In this scenario, the Cox proportional hazards model could be applied. In the situations created, imputation showed adequate results, including in the performance analyses. The technique had a satisfactory effectiveness when the missing mechanism was MAR (Missing At Random) and the percent of missing data was 10. Imputing the missing information and combining the estimates of the 10 resulting datasets produced a bias of 0,0011 to black women and 0,0015 to brown (mixed-race) women, what corroborates the efficiency of multiple imputation in this scenario. In the second paper, a tutorial was created to guide the application of multiple imputation in epidemiologic studies, which should facilitate the use of the technique by Brazilian researchers who are still not familiarized with the procedure. Basic steps and important decisions necessary to impute a dataset are presented and one of the scenarios of the first paper is used as an application example. All the analyses were performed at R statistical software, version 2.15 and the scripts are presented at the end of the text.

Keywords: Missing data. Multiple imputation. Survival analysis. Tutorial

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BLR	*Bayesian Linear Regression* – Regressão Linear Bayesiana
DF	Dados Faltantes
DMQ	Desvio Médio Quadrático
EPS	Estudo Pró-Saúde
FCS	*Fully Conditional Specification* – Especificação Condicional Completa
IBGE	Instituto Brasileiro de Geografia e Estatística
IM	Imputação múltipla
IPW	*Inverse Probability Weighting* – Ponderação pela Probabilidade Inversa
JM	*Joint Modelling* – Modelagem Conjunta
MAR	*Missing At Random* – Dado faltante aleatório
MCAR	*Missing Completely At Random* - Dado faltante completamente aleatório
MCMC	*Markov Chain Monte Carlo* – Cadeia de Markov-Monte Carlo
MICE	*Multiple Imputation by Chained Equations* – IM por equações em cadeia
MNAR	*Missing Not At Random* – Dado faltante não aleatório
MU	Mioma(s) Uterino(s)
MV	Máxima Verossimilhança
PMM	*Predictive Mean Matching* – Pareamento pela Média Predita
RH	Razão de Hazards

SUMÁRIO

APRESENTAÇÃO

Este projeto está inserido no Estudo Pró-Saúde (EPS), um programa de investigação que tem como população-alvo funcionários técnico-administrativos de uma universidade no estado do Rio de Janeiro. Entre seus objetivos figuram a descrição de perfis de morbidade e seus fatores de risco, de comportamentos ligados à saúde e a investigação de determinantes sociais associados a tais perfis, em um contexto urbano no Brasil. O estudo também busca contribuir na promoção da saúde e prevenção de doenças em sua população-alvo.

O projeto existe desde 1999 e, desde então, cerca de 4000 funcionários técnico-administrativos vêm sendo acompanhados periodicamente. As atividades do grupo de pesquisa estão vinculadas, sobretudo, aos programas de pós-graduação do Instituto de Medicina Social e da Escola Nacional de Saúde Pública Sérgio Arouca (ENSP/FIOCRUZ), mas passaram a incorporar, nos últimos anos, pesquisadores oriundos da Universidade Federal do Rio de Janeiro (UFRJ), Universidade Federal do Rio Grande do Sul (UFRGS), e da Universidade da Califórnia/ Berkeley (UC/B).

O Pró-Saúde/UERJ desenvolve suas atividades prioritariamente na área de pesquisas epidemiológicas que utilizam métodos quantitativos e é exatamente esta a vertente na qual este projeto está inserido. Nos últimos anos, com o avanço nas pesquisas no Brasil e no mundo, técnicas mais elaboradas e complexas foram criadas tanto no âmbito epidemiológico quanto bioestatístico. Com isso, os estudos realizados no presente precisam incorporar tais técnicas em sua metodologia.

Desde o início da década passada, o Pró-Saúde tem demonstrado interesse em métodos estatísticos mais complexos, incluindo nestes a imputação múltipla (Moreno, 2004). Porém, a aproximação com as técnicas de imputação se deu de forma inicial e não foi estendida a outras análises do Pró-Saúde. A vontade de explorar técnicas de imputação múltipla advém da necessidade de se tratar a não-resposta, tão comum em pesquisas epidemiológicas. O programa, apesar dos imensos esforços em evitar a ocorrência de dados faltantes (Faerstein *et al.*, 2005), também sofre do problema e necessita de estratégias que auxiliem a lidar melhor com essas questões.

No Brasil, infelizmente, ainda são poucos os programas de investigação que resolvem lidar adequadamente com o problema e que se dedicam ao uso de métodos avançados de imputação de dados. Desta maneira, com esta dissertação, pretende-se aprofundar os avanços realizados no Pró-Saúde na área de imputação múltipla, criando uma ferramenta que facilite o uso da técnica não só neste estudo, mas que seja passível de uso por qualquer grupo de

pesquisa cujos dados tenham estrutura semelhante aos do EPS. Espera-se contribuir para divulgação da técnica ainda incipiente e pouco explorada no Brasil, especificamente na área de análise de sobrevivência, na qual o uso da imputação parece não estar totalmente consolidado inclusive no cenário internacional.

Para elaboração desta dissertação, além de se utilizar as referências clássicas sobre o tema, buscou-se a leitura dos trabalhos mais recentes – metodológicos ou que envolvessem aplicação prática de métodos de imputação - e o acompanhamento semanal de publicações da área. Como palavras-chave na busca por artigos do tema no PubMed, utilizou-se inicialmente *missing data* e *(multiple) imputation* e, na sequência, a busca foi filtrada por *epidemiology*. Para acompanhar publicações específicas de imputação múltipla em estudos de sobrevivência, também se utilizou a palavra-chave *survival* e, para acompanhar publicações de estudos longitudinais, a palavra-chave *longitudinal* foi empregada.

Este documento apresenta a aplicação da imputação múltipla em estudos de sobrevivência e está organizado em 8 seções. A primeira seção (Revisão da literatura) oferece um panorama do tema em questão, apresentando inicialmente os conceitos (subseções 1.1 e 1.2), métodos para lidar com dados faltantes (subseção 1.2), divididos entre as técnicas *ad hoc* (subseção 1.2.1), técnicas baseadas em verossimilhança (subseção 1.2.2) e técnicas de imputação múltipla (subseção 1.2.3). A segunda seção oferece as justificativas para a realização deste estudo. Em seguida são apresentados os objetivos desta dissertação (seção 3). A quarta seção descreve os métodos utilizados e traz uma breve descrição das bases metodológicas aqui empregadas (subseção 4.1). Na seção seguinte, os resultados deste estudo são apresentados na forma de dois artigos: o primeiro deles, intitulado “Imputação múltipla de dados faltantes em análise de sobrevivência – aplicação no Estudo Pró-Saúde”, a ser submetido ao periódico BMC - Medical Research Methodology; e, o segundo, intitulado “Imputação múltipla de dados – tutorial para aplicação em estudos epidemiológicos”, a ser submetido ao periódico Cadernos de Saúde Pública. Ao final, são apresentadas algumas considerações (seção 6), as referências utilizadas para elaboração desta dissertação (seção 7) e, nos apêndices (seção 8), uma breve descrição da população avaliada (subseção 8.1) e os scripts utilizados nas análises de dados (subseção 8.2).

1 **REVISÃO DA LITERATURA**

1.1 **Considerações gerais e conceitos básicos**

Dados faltantes são um problema comum em estudos epidemiológicos, e, dependendo da forma como ocorrem, as estimativas obtidas a partir da análise dos dados existentes podem não ser válidas (Rothman *et al.*, 2008). A ocorrência dos dados faltantes em um pequeno percentual de algumas variáveis pode resultar em um grande número de observações com alguma informação não disponível (Horton & Kleinman, 2007) e, de acordo com He (2010), as razões para esta indisponibilidade são diversas: vão desde o preenchimento inadequado de algum item por parte do entrevistado ou a recusa em responder determinada questão até a perda do indivíduo ao longo do estudo, no caso de estudos longitudinais (Nunes *et al.*, 2010). Ignorar a não-resposta pode ser problemático, já que a inferência a ser feita a partir dos dados disponíveis pode estar comprometida (Haukoos & Newgard, 2007).

Existem diversas maneiras de se lidar com os dados faltantes (*missing data*). Estas variam desde o descarte de indivíduos com alguma informação faltante – a chamada análise de observações completas - a métodos estatísticos mais complexos, que envolvem o "resgate" do que está faltando com base nas informações disponíveis (Donders *et al.*, 2006). Little e Rubin (1987) propõem uma classificação quanto ao procedimento de análise utilizada nestas situações, a saber: métodos baseados na análise de observações completas, baseados em imputação, em atribuição de pesos ou baseados em modelos. Segundo os autores, os métodos baseados na análise de observações completas envolvem o descarte das unidades com alguma informação faltante – procedimento que pode levar a vieses e que, em geral, não é eficiente. Dentre os métodos baseados em imputação, estão tanto aqueles considerados de imputação única e os de imputação múltipla. Entre as técnicas que envolvem a atribuição de pesos figuram a Ponderação pela Probabilidade Inversa (do inglês, *Inverse Probability Weighting* - IPW) (Seaman & White, 2011). Já os métodos baseados em modelos dizem respeito à aplicação de técnicas de máxima verossimilhança, cujas vantagens são a flexibilidade e o fato de não se necessitar de métodos de imputação *ad hoc* (Little & Rubin, 1987).

Não existe uma solução ideal para se lidar com qualquer problema de dados faltantes. Alguns autores (Harel & Zhou, 2007; Buhi *et al.*, 2008) afirmam que mesmo um pequeno percentual de não-resposta deve ser avaliado com cautela e, além disso, o pesquisador deve escolher o método que seja capaz de maximizar a acurácia e precisão das estimativas.

Muitos destes métodos disponíveis – sobretudo os de imputação - envolvem a adoção de pressupostos quanto ao mecanismo de ocorrência dos dados faltantes e quanto ao padrão desta ocorrência. Quanto aos mecanismos, estes podem ser: dado faltante completamente aleatório (em inglês, *missing completely at random* – MCAR), dado faltante aleatório (*missing at random* – MAR) e dado faltante não aleatório (*missing not at random* – MNAR).

Dados faltantes completamente aleatórios (MCAR) ocorrem quando a informação faltante é independente da variável de exposição e do desfecho. Neste caso, os dados observados constituem uma subamostra aleatória dos dados totais (Little & Rubin, 1987) – em geral, esse pressuposto não costuma ser a realidade dos estudos epidemiológicos (Grittner *et al.*, 2011).

Já o dado faltante aleatório (MAR) ocorre quando a informação faltante depende apenas do que foi observado e não do que não foi, e, tal informação pode ser reconstituída com base nos dados disponíveis. Como afirmam Grittner e colaboradores (2011), o dado faltante pode diferir entre os subgrupos de uma amostra, mas é aleatório dentro de cada um deles. Tanto MAR quanto MCAR são considerados "dados faltantes ignoráveis": quando ocorrem, pode-se ignorar as razões que explicam porque o dado não foi observado e empregar uma técnica apropriada para lidar com o problema (Buhi *et al.*, 2008). Com base no pressuposto de os dados faltantes serem do tipo MAR, grande parte das técnicas de imputação foi desenvolvida (Rubin, 1987; Donders *et al.*, 2006).

Dado faltante não aleatório (MNAR) acontece quando a informação faltante depende do que não foi observado e, possivelmente, daquilo que foi de fato observado. É também chamado "dado faltante não ignorável" (Rubin, 1987), e os modelos de análise comumente utilizados não são capazes de lidar adequadamente com os dados incompletos resultantes deste mecanismo (Buhi *et al.*, 2008). Segundo Nunes (2007), dados mais sujeitos a serem MNAR são aqueles encontrados nos extremos da distribuição, com valores maiores ou menores do que o padrão observado na amostra.

Um exemplo hipotético destes mecanismos no âmbito do estudo Pró-Saúde diz respeito à pergunta sobre realização de exame de mama pelas mulheres. Caso a mulher deixe de responder por ter faltado ao trabalho, ou estar de férias ou de licença, o dado faltante poderia ser completamente aleatório (MCAR). Poderia ser não aleatório (MNAR) caso a mulher deixe de responder a pergunta exatamente por não realizar o exame rotineiramente. E, caso a resposta permanecesse em branco por razões não relacionadas à não realização do exame, mas a outras questões presentes no questionário (por exemplo, à renda ou à escolaridade), o dado faltante poderia ser considerado aleatório (MAR).

Ainda em relação aos mecanismos de ocorrência de dados faltantes, determinar se este é do tipo MAR ou MNAR depende dos dados e da plausibilidade em se assumir um mecanismo ou outro, já que a adoção deste pressuposto não é testável (Little & Rubin, 1987). Diversos autores afirmam que, caso o dado faltante seja inevitável em determinado estudo, pode-se fazer com que ele se aproxime o máximo possível do mecanismo MAR e se afaste de MNAR através da inclusão de medidas adicionais durante o planejamento do estudo, o que faz com que a probabilidade do dado faltante ser MAR aumente (Harel & Zhou, 2007; Buhi *et al.*, 2008).

Quanto ao padrão de ocorrência dos dados faltantes, de maneira geral, este pode ser monotônico ou não-monotônico (figura 1). No primeiro, comum em estudos longitudinais, blocos de variáveis apresentam cada vez mais dados faltantes, ou de acordo com as ondas de seguimento ou de acordo com a sequência de variáveis em uma mesma onda. No segundo, tal padrão crescente de não-resposta não é observado (Little & Rubin, 1987). Há métodos específicos para lidar com um ou outro (Haukoos & Newgard, 2007), e os modelos utilizados para padrões monotônicos são considerados mais simples (Horton & Kleinman, 2007). Cabe ressaltar que a não-resposta pode ocorrer em uma unidade inteira – quando o sujeito selecionado se recusa a continuar no estudo – ou em alguns itens – nos quais os indivíduos se negam a responder algumas perguntas por constrangimento, por exemplo.

Nos últimos anos, alguns autores (Kristman *et al.*, 2005; Donders *et al.*, 2006; Buhi *et al.*, 2008; Wang *et al.*, 2011) buscaram avaliar diferentes métodos para lidar com o problema dos dados faltantes aplicando-os a diversos bancos de dados e comparando-os, tanto em estudos longitudinais como seccionais. Outros autores (Haukoos & Newgard, 2007; Klebanoff & Cole, 2008; Seaman & White, 2011) se dedicaram a revisar a literatura acerca do problema, a fim de identificar as técnicas mais comumente utilizadas. A seguir, alguns dos métodos existentes serão apresentados.

1.2 **Técnicas de análise estatística com dados faltantes**

Para lidar com os dados faltantes, pode-se decidir entre o descarte das informações não observadas – técnica mais comumente utilizada em Epidemiologia (Nunes *et al.*, 2009); as técnicas de imputação única – e, entre estas, um método que parece ser consensualmente não recomendado é a imputação pela média ou pela mediana (Sinharay *et al.*, 2001); ou as de imputação múltipla. A seguir, uma breve descrição de algumas destas técnicas, com ênfase

especial na imputação múltipla. Dentre as técnicas *ad hoc*, menciona-se as de exclusão dos indivíduos com alguma porção de informação não observada e as de imputação única. Na sequência, são apresentadas algumas técnicas baseadas em verossimilhança e os conceitos que subjazem à imputação múltipla.

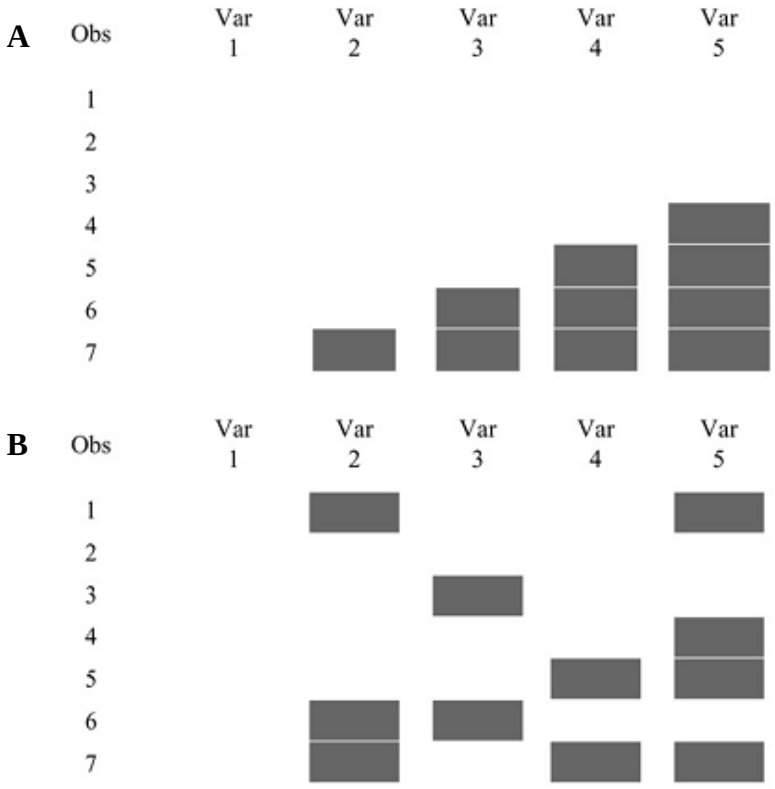

Figura 1 – Padrões de ocorrência de dados faltantes. Em A, observa-se um padrão monotônico; em B, um padrão não-monotônico. Note que as colunas podem se referir a diferentes variáveis em uma mesma onda ou à mesma variável em diferentes ondas de um estudo longitudinal (adaptado de Haukoos & Newgard, 2007).

1.2.1 Técnicas *ad hoc*

De maneira geral, as técnicas *ad hoc* para se lidar com os dados faltantes são: análise de observações completas, análise de observações disponíveis, análise de observações completas com ponderação, *inverse probability weighting* (IPW), imputação única (imputação

pela média ou mediana, imputação por regressão, por regressão estocástica, imputação via *hot* ou *cold deck*, último valor observado).

Em uma análise de observações completas (do inglês, *Complete- Case analysis* ou *Listwise Deletion*), apenas as observações com a totalidade de informação nas variáveis a serem estudadas são considerados na análise (Sinharay *et al.,* 2001) – em outras palavras, ocorre exclusão de todos os indivíduos que tenham dados faltantes em qualquer variável. Se um participante do estudo deixa de responder a alguma questão, toda informação referente a esse indivíduo é excluída. Esta exclusão costuma ser o *default* da maioria dos aplicativos de análise estatística (Moons *et al.,* 2006; He, 2010). Por conta disso, ocorre redução do poder estatístico (pela redução do tamanho amostral), ou seja, ocorre aumento do erro tipo II. Também pode ocorrer viés, pois os indivíduos com dados completos podem diferir substancialmente daqueles excluídos da análise – e, isto seria potencialmente problemático em estudos que envolvem doenças raras, já que o grupo a ser descartado pode conter uma grande proporção dos participantes que possuem a doença em questão (Schafer & Graham, 2002). Com isso, a inferência a partir desta análise pode estar comprometida (Buhi *et al.*,2008). Se o dado faltante é do tipo MCAR, alguns autores (Haukoos & Newgard, 2007; He, 2010) mencionam que a análise de observações completas pode não resultar em viés.

Semelhante à abordagem anterior, existe a chamada análise de observações disponíveis (*Available case analysis* ou *Pairwise Deletion*), na qual se limita a análise aos indivíduos com informações completas em cada variável (ou grupo de variáveis) sendo analisada por vez, isto é, haverá variação no número de indivíduos disponível para cada análise a ser realizada (Haukoos & Newgard, 2007). Apesar de parecer mais eficiente do que a análise de observações completas, pode levar aos mesmos problemas e, além disto, as estimativas de cada análise podem ser baseadas em diferentes conjuntos de indivíduos (Sinharay *et al.,* 2001), por isso, muitos analistas já optam pela análise de observações completas e excluem todos os indivíduos com alguma informação faltante no início das análises.

Uma terceira abordagem citada como possível saída para minimizar os problemas da análise de observações completos é mencionada por Haukoos e Newgard (2007). Pode ser denominada análise de observações completas com ponderação – uma modificação da análise de casos completos que pondera observações completas e incompletas de maneira diferente, a fim de considerar potenciais vieses, e que pode ser utilizada em dados de inquéritos. Ponderar as observações desta maneira permite ajustar por potenciais vieses introduzidos por indivíduos que não participam do estudo, mas seriam elegíveis para o mesmo (Little & Rubin,

1987). Apesar de reduzir os vieses em análises subsequentes, ocasiona redução da precisão das estimativas e poderia ser melhor aplicada em bancos de dados maiores, nos quais a perda de precisão não é tão importante (Haukoos & Newgard, 2007).

Uma maneira mais específica de se ponderar as informações é através da chamada Ponderação pela Probabilidade Inversa - *Inverse Probability Weighting* (IPW). O método consiste em ponderar as informações pelo inverso da probabilidade de terem sido observadas. Com isso, apenas as observações completas são incluídas na análise, mas os pesos são utilizados para reequilibrar o conjunto de dados, de forma que se tornem representativos de toda a amostra. IPW também pode ser utilizada para ajuste por diferentes frações amostrais em um inquérito, de forma a reequilibrar a amostra e torná-la representativa da população (Seaman & White, 2011). Estes métodos de ponderação podem se tornar intratáveis quando há dados faltantes em muitas variáveis, particularmente quando o padrão de ocorrência é não-monotônico (Horton & Kleinman, 2007). Nos últimos anos, têm se buscado utilizar IPW em conjunto com a imputação múltipla, a fim de solucionar limitações de ambos os métodos (Seaman *et al.*, 2012).

O que se denomina "imputação única" compreende um conjunto de técnicas nas quais os valores faltantes são substituídos por um único valor plausível, que, na teoria, seria aquele que deveria ter sido observado (Haukoos & Newgard, 2007; Nunes, 2007; He, 2010). Em seguida, são conduzidas as análises estatísticas de rotina, em um banco de dados completo, assumindo que toda a informação foi observada.

O método mais simples de imputação única é a imputação pela média ou pela mediana, que consiste em substituir todos os dados faltantes pela média ou mediana da variável em questão. Porém, mesmo permitindo a inclusão de todas as observações ao final e se tratando de uma técnica de fácil execução, os resultados gerados não são válidos, exceto se o dado faltante for do tipo MCAR (Sinharay *et al.*, 2001). Os principais problemas decorrentes do uso da média/mediana para imputar são o fato de que a variância e covariância estarão subestimadas, o que leva a atenuação das correlações entre as variáveis e mudança na distribuição da variável sendo imputada (Buhi *et al.*, 2008). Além disto, tal subestimação aumentará conforme a proporção de dados faltantes for maior, fazendo com que a precisão das estimativas aumente de maneira incorreta, juntamente com a probabilidade de ocorrência de erro tipo I (Haukoos & Newgard, 2007).

Outro método de imputação única utilizado em Epidemiologia é a chamada imputação por regressão, na qual o valor faltante de uma variável é substituído por um único valor predito por um modelo de regressão baseado nos dados observados (Sinharay *et al.*, 2001).

Pode ser capaz de gerar aproximações razoáveis, dependendo do mecanismo de ocorrência dos dados faltantes, contudo, indivíduos que possuem os mesmos valores nas mesmas covariáveis terão valor predito idêntico (Nunes, 2007), e, haverá subestimação da variância porque o método assume que não há nenhum erro residual ao redor da reta de regressão (Sinharay *et al.,* 2001). O método vem sendo utilizado em Epidemiologia em estudos que desejam imputar determinado desfecho em saúde em sujeitos perdidos durante um estudo de seguimento (Kristman *et al.,* 2005), porém, sabe-se que a imputação do desfecho se trata de uma prática pouco recomendada.

Uma alternativa é a imputação por regressão com componente estocástico, na qual o valor faltante é substituído pelo valor predito somado a um termo de erro aleatório (Sinharay *et al.*, 2001; Haukoos & Newgard, 2007). É capaz de incorporar incerteza ao valor imputado, resolvendo em parte, a principal desvantagem relacionada à variância da imputação por regressão. Entretanto, a exemplo do método anterior, ainda falha em considerar a incerteza (ou seja, um componente adicional da variância) inerente ao processo de imputação (Haukoos & Newgard, 2007).

Outro método que vem recebendo destaque nos últimos anos é o de *hot deck* – nele, cada valor faltante é substituído por um valor observado selecionado entre os indivíduos considerados semelhantes àquele cuja informação não foi observada – chamados "doadores". Esta seleção pode ser através amostragem com reposição entre os possíveis "doadores" ou através do método do vizinho mais próximo – onde algum critério é determinado para selecionar aquele com características mais próximas do indivíduo a ser imputado (Nunes, 2007; Grittner *et al.,* 2011). O doador é alguém que compartilha com o "receptor" os mesmos padrões de resposta para um determinado conjunto de variáveis, definido anteriormente pelo pesquisador (Buhi *et al.,* 2008). O procedimento de substituição das informações faltantes é realizado até que todo banco de dados esteja completo. Dentre as vantagens do método pode-se mencionar sua simplicidade; o fato de não requerer a adoção de pressupostos quanto à distribuição de parâmetros; e, os valores imputados podem refletir melhor a distribuição daquela variável e não estarão fora dos domínios de variação da mesma (Grittner *et al.*, 2011). O *hot deck* também preserva as distribuições marginais, mas pode distorcer relações entre as variáveis (Sinharay *et al.,* 2001). Além disto, não existe um critério bem definido para guiar a seleção do que pode ser considerado "semelhante" e o método assume que respondentes e não respondentes são idênticos, o que nem sempre é plausível (Patrician, 2002; Haukoos & Newgard, 2007). Vale ressaltar que existe uma variação do mesmo que faz parte das técnicas de imputação múltipla (Wang *et al.*, 2011).

Com estratégia semelhante ao *hot deck*, há o denominado *cold deck*, no qual o valor semelhante não é selecionado a partir do banco de dados existente, e sim de uma fonte de dados externa, independente do banco sendo utilizado – por exemplo, uma média populacional (Grittner *et al.*, 2011). Uma desvantagem adicional do *cold deck* em relação ao *hot deck* é que as informações provenientes dessa fonte externa podem diferir das informações disponíveis no banco de dados de análise, adicionando mais um componente de viés à estimação dos parâmetros. Em geral, é um método pouco utilizado e não recomendado (Haukoos & Newgard, 2007).

Um método aplicado em estudos longitudinais que sofrem com perda de seguimento é o do último valor observado (do inglês, *Last observation carried forward* (LOCF) – nele, utiliza-se a última observação realizada do indivíduo para substituir o dado faltante (Haukoos & Newgard, 2007; Grittner *et al.*, 2011), isto é, se um indivíduo participa da 1ª onda de um estudo de seguimento, não participa da 2ª e participa das demais, utiliza-se a informação da 1ª onda para completar a 2ª. Da mesma forma, o método conhecido como *Next observation carried backward* consiste em substituir informação faltante de uma fase pelo dado obtido em uma onda posterior (Nunes, 2007) - utilizando o exemplo anterior, se o indivíduo participa da 1ª onda, não participa da 2ª e participa da 3ª, a informação da 3ª fase é utilizada para completar o dado faltante na 2ª. As duas técnicas têm por vantagem o fato de serem simples e não exigirem suposições quanto à distribuição das variáveis ou a determinação de preditores do dado faltante (Grittner *et al.*, 2011). A principal crítica é que se baseiam no pressuposto de que a informação do indivíduo permanece inalterada entre duas ondas, o que geralmente não é plausível (Haukoos & Newgard, 2007). Grittner e colaboradores (2011) também mencionam que os métodos podem levar a estimativas incorretas da variância.

Ainda em estudos longitudinais, Nunes (2007) menciona outras abordagens semelhantes às anteriores: usar a média/mediana dos valores conhecidos do indivíduo prévios à fase em que há dados faltantes; usar a média ou mediana de todos os valores (anteriores e posteriores) ao dado faltante para cada indivíduo ou observação; utilizar a média do último e do próximo valor observado apenas. Alguns destes métodos são também utilizados por Engels e Diehr (2003).

Os métodos de análise de observações completas, de ponderação e de imputação única antes mencionados - apesar da sua fácil execução - podem levar a resultados enviesados e, no caso daqueles considerados de imputação única, por tratar os valores imputados como verdadeiros, resultarão em superestimação da precisão das estimativas das associações do estudo – por subestimarem o erro-padrão, levando a intervalos de confiança mais estreitos

(Patrician, 2002; Donders *et al.*, 2006). Porém, como afirmado anteriormente, não existe um método de imputação ideal e a decisão de imputar ou não e qual método utilizar depende das características dos dados que se tem, dos recursos disponíveis, e, da plausibilidade da adoção de determinados pressupostos necessários a todos os métodos.

1.2.2 Técnicas baseadas em verossimilhança

Máxima verossimilhança (MV) se refere a uma abordagem estatística muito utilizada para estimação de parâmetros, como por exemplo, coeficientes lineares da regressão logística. Além das abordagens nas quais a MV é comumente utilizada, sabe-se que ela pode ser adaptada para lidar com o problema dos dados faltantes (Allison, 2001). O princípio básico da MV é selecionar como estimativas dos parâmetros os valores que, se verdadeiros, são capazes de maximizar a probabilidade de se obter a amostra que foi observada. Os estimadores de MV possuem algumas propriedades importantes: sob diversas condições, são consistentes, assintoticamente eficientes e assintoticamente normais.

Quando a estimação envolve dados faltantes, se o mecanismo de não-resposta for ignorável, pode-se estimar os parâmetros através de técnicas que envolvem MV e obter estimativas apropriadas, sobretudo se o padrão de ocorrência de dados faltantes for monotônico (Allison, 2001). As estimativas resultantes são mais eficientes porque todos os dados observados são utilizados e não necessariamente ocorre o preenchimento dos valores faltantes. Alguns autores mencionam que, em determinadas situações, tais métodos podem ser mais eficientes do que os de imputação múltipla de dados (Graham *et al.*, 2007).

Os métodos de MV são específicos para o problema que se está lidando e geralmente envolvem modelos e ferramentas computacionais mais complexas – percentuais maiores de dados faltantes resultarão em taxas de convergência menores (Newgard & Haukoos, 2007).

O método de MV bastante utilizado para se lidar com os dados faltantes é o chamado algoritmo EM, aplicado quando se deseja obter a estimativa de um parâmetro a partir de um conjunto de dados incompleto (Nunes, 2007). A sigla EM se refere aos dois passos necessários à técnica: *Expectation* e *Maximization*. No passo E, estima-se os dados faltantes para completar a matriz de dados, de maneira similar à imputação por regressão. No passo M, a partir do banco de dados completos, utiliza-se os dados imputados em E para atualizar o vetor de médias e a matriz de covariância, que são usados para atualizar as estimativas dos dados faltantes no passo E. Os dois passos são repetidos várias vezes em um processo

iterativo até a convergência para a estimativa de MV (Allison, 2001; Patrician, 2002; Nunes, 2007).

Enders (2010) traz uma descrição detalhada de outros métodos que utilizam MV e exemplos úteis de sua aplicação.

1.2.3 Técnicas de Imputação Múltipla

1.2.3.1 Considerações gerais

Apesar das muitas técnicas anteriormente apresentadas, a literatura aponta a imputação múltipla (IM) como a mais adequada em diversas situações de não-resposta (Zhou *et al.*, 2001; Klebanoff & Cole, 2008). A IM surge como uma alternativa para gerar estimativas válidas e que considerem a incerteza devido à imputação – fato desconsiderado na imputação única. Isto é possível por meio da criação de alguns bancos de dados imputados e pela posterior combinação dos resultados provenientes das análises destes (Sterne *et al.*, 2009).

A técnica proposta por Rubin na década de 1970 (Rubin, 1976; Rubin, 1978) possui características bayesianas, já que não se deseja imputar um valor único, mas uma distribuição preditiva dos valores faltantes dado aquilo que foi observado. Basicamente, consiste em criar múltiplas cópias do banco de dados (‘*m*’ bancos, sendo $m>1$), com os dados faltantes substituídos por valores imputados. Para gerar tais valores, são usados os dados observados. Sequencialmente, podem ser utilizadas técnicas estatísticas tradicionais e o modelo de interesse é aplicado a cada banco gerado. As “regras de Rubin” são então aplicadas para calcular os erros-padrão das estimativas, considerando a variabilidade dos resultados nos diferentes bancos imputados. Desta forma, a incerteza associada à imputação é levada em conta (Canizares *et al.*, 2004; Sterne *et al.*, 2009). De acordo com He (2010), se a imputação é conduzida de maneira plausível, ela fornecerá predições razoáveis sobre os dados faltantes e a variabilidade entre as imputações deve refletir um grau de incerteza apropriado.

Do ponto de vista inferencial, uma das principais razões de se utilizar a IM é o fato de ela poder incorporar informações observadas e não observadas (Harel & Zhou, 2007) – já que o principal pressuposto necessário à imputação é de que os dados faltantes são do tipo MAR e podem ser imputados com base no que foi efetivamente observado. Com isso, a IM pode ser utilizada quando a não-resposta ocorre em um ou mais itens de um indivíduo e não na unidade

toda – ou seja, requer que os dados daquele indivíduo com algum dado faltante possuam alguma porção de informação observada (Haukoos & Newgard, 2007).

A IM compartilha com a imputação simples algumas características, tais como a possibilidade de se usar métodos de análise para bancos completos e a habilidade de incluir o conhecimento do pesquisador (Nunes, 2007). Além disto, a IM possui vantagens adicionais: quando as imputações são aleatoriamente realizadas, em uma tentativa de representar a distribuição dos dados, a IM é capaz de aumentar a eficiência da estimação; quando as imputações (os "*m*" bancos de dados) representam realizações aleatórias repetidas sob um mesmo modelo de não-resposta, inferências válidas que reflitam a variabilidade adicional por conta dos valores faltantes daquele modelo são facilmente obtidas pela combinação das inferências nos bancos completos; se as imputações são geradas a partir de diferentes modelos, pode-se estudar a sensibilidade das inferências para vários modelos de não-resposta simplesmente usando métodos para dados completos repetidamente (Rubin, 1987).

Rubin (1987), em seu livro clássico sobre o tema, ainda menciona três desvantagens principais da IM, quando comparada à imputação simples, a saber: primeiro, o trabalho necessário para se realizar a IM é muito maior – o que Nunes (2007) considera um pequeno esforço, dados os ganhos na inferência que a IM pode proporcionar; segundo, mais espaço é necessário para armazenar os conjuntos de bancos da imputação múltipla; terceiro, o trabalho de análise do banco imputado via IM é maior do que o necessário para o banco imputado de forma única. Os dois últimos problemas mencionados podem ser considerados de menor importância atualmente, sobretudo devido à grande quantidade de ferramentas computacionais disponíveis para imputação múltipla. Horton e Kleinman (2007) realizaram uma revisão sobre métodos e aplicativos disponíveis para IM e comentam e existência de pelo menos dez pacotes aplicáveis a diversos programas estatísticos já conhecidos. Com isso, e com o aumento da capacidade dos computadores, o esforço necessário à IM passou a ser menor.

Diversos autores vêm aplicando a imputação múltipla em seus trabalhos (Kmetic *et al.*, 2002; Burns *et al.*, 2011; Mumford *et al.*, 2011), por reconhecerem que a técnica é superior às demais quando aplicada a diversos problemas de dados faltantes. Klebanoff e Cole (2008) trazem uma revisão do uso da IM em Epidemiologia e ressaltam que, se os métodos de imputação forem adequadamente aplicados, são capazes de reduzir a presença de viés e de imprecisão nos estudos conduzidos na rotina dos epidemiologistas. Nunes (2007) também faz um breve levantamento de algumas aplicações da IM em Epidemiologia.

Apesar de se conhecer os problemas relacionados à não-resposta e às questões quanto ao comprometimento da validade das inferências que emergem dos dados faltantes e do uso de técnicas *ad hoc* para reconstituição da informação não observada, a aplicação da imputação múltipla na análise de estudos epidemiológicos ainda é incipiente no Brasil. Em recente revisão realizada nas bases de dados nacionais (LILACS e SCIELO), utilizando como palavras-chave "dados faltantes" e "imputação (múltipla)", e em bases de dados internacionais (PubMed/MedLine), utilizando as palavras "*missing data*", "*(multiple) imputation*" e "*Brazil*", apenas três publicações que tratam da questão dos dados faltantes aplicados à Epidemiologia foram encontradas (Nunes *et al.,* 2009; Nunes *et al.,* 2010; Camargos *et al.,* 2011).

Nunes e colaboradores (2009) conduziram um estudo de simulação com dados provenientes de pacientes internados no Hospital de Clínicas de Porto Alegre (Porto Alegre, RS). Utilizaram um modelo de regressão logística, simularam os dados faltantes a partir de um conjunto de dados considerado completo, e imputaram a variável albumina. A imputação foi conduzida no *software* R, através do pacote MICE, utilizando as abordagens de *Predictive Mean Matching* (PMM) e *Bayesian Linear Regression* (BLR). As autoras compararam dois percentuais distintos de dados faltantes (5 e 20%) e utilizaram o banco completo como referência a fim de verificar a efetividade do procedimento de IM. Como conclusões, recomendam a imputação como forma de aumentar a confiabilidade dos resultados e o poder estatístico das análises, por conta do aumento efetivo do conjunto de dados e, além disso, mencionam a superioridade da IM sobre a imputação simples e ressaltam a necessidade de mais trabalhos metodológicos sobre seu uso.

Em trabalho publicado em 2010, as mesmas autoras (Nunes *et al.,* 2010) utilizando o mesmo conjunto de dados e a variável albumina, compararam métodos de imputação múltipla – através de BLR - e única – imputação pela mediana e pelo valor do limite inferior da faixa de normalidade (denominado "método do valor normal"). Dentre os modelos de imputação múltipla, consideraram modelos que incluíam ou não o desfecho de interesse, em regressões logísticas multivariadas. Novamente, utilizaram o *software* R e o pacote MICE. Como resultados, comentam que mesmo a imputação simples apresenta melhores resultados do que a análise de observações completas e ressaltam as qualidades e a superioridade da IM em relação à imputação única. Mais uma vez, afirmam que as técnicas de imputação têm recebido pouca atenção pelos pesquisadores e apontam a carência de textos metodológicos para que os pesquisadores utilizem a IM com maior confiança.

Em 2011, utilizando dados de um inquérito de saúde domiciliar (*Saúde em Beagá*), Camargos e colaboradores (2011), também utilizaram regressão logística multivariada para imputar dados de Índice de Massa Corporal (IMC) aferido. Para gerar os dados faltantes, partiram da informação faltante de IMC relatado e, nos indivíduos que possuíam essa informação nos dados de IMC aferido, realizaram a exclusão do dado. Com isso, e assumindo que o IMC relatado e o aferido eram semelhantes, obtiveram um banco de dados incompleto. Compararam os resultados do IMC aferido no banco completo com resultados após a exclusão nesta variável, utilizando a análise de observações completas e IM (não mencionam a abordagem utilizada). Utilizaram o *software* Stata, versão 11. Novamente, o banco imputado demonstrou resultados mais satisfatórios nas análises. Os autores também reforçam a necessidade de mais trabalhos para acumulação de evidências no tema.

Pode-se observar que a IM no Brasil se trata de um tema recente, ainda pouco explorado e cuja metodologia requer maior apreciação por parte dos pesquisadores no país. O que não se sabe é se realmente as técnicas de imputação não são utilizadas no Brasil, ou se apenas não são mencionadas quando das análises dos diferentes estudos aqui publicados. A questão que emerge é que a técnica é vastamente utilizada em grandes pesquisas no exterior (Moons *et al.*, 2006; Burns *et al.*, 2011; Grittner *et al.*, 2011; Mumford *et al.*, 2011) e ainda parece não ser largamente aplicada no Brasil.

Segundo Nunes e colaboradores (2010), a razão para o pouco (ou não) uso poderia ser a exigência de técnicas computacionais mais elaboradas, fato que já não serve como justificativa, já que diversos programas estatísticos, sejam eles pagos ou de domínio público, dispõem de recursos para a aplicação de estratégias de imputação múltipla (Horton & Kleinman, 2007). Outra possível razão levantada pelas autoras seria a complexidade das suposições requeridas para uso destes métodos.

1.2.3.2 Procedimento geral de Imputação Múltipla

Rubin (1987) recomenda que a imputação seja criada por meio de mecanismos bayesianos, através de: especificação de um modelo paramétrico para os dados completos sob MAR; suposição de uma distribuição *a priori* (não informativa) para os parâmetros desconhecidos do modelo; e, simulação de múltiplas amostras independentes da distribuição *a posteriori* condicional dos valores faltantes, dada a informação observada, pelo teorema de Bayes (Enders, 2010; He, 2010). Em geral, os modelos de imputação seguirão este formato.

Supondo Y = (Y_1,Y_2,...,Y_k) como um vetor de *k* variáveis aleatórias com uma distribuição *k*-variada P(Y|θ). Assume-se que a distribuição conjunta multivariada de Y é completamente especificada por θ, um vetor de parâmetros desconhecidos. Seja a matriz y = (y_1, ... , y_n) com y_i = (y_{i1},y_{i2}, ... , y_{ik}), *i* =1, ..., *n* uma amostra independente e identicamente distribuída (i.i.d.) do vetor Y. A matriz y é parcialmente observada, no sentido em que cada coluna em y tem um valor faltante. De maneira geral, o procedimento padrão para se gerar múltiplas imputações y* a partir de y^{obs} pode ser resumido em três passos:

1. Calcular a distribuição *a posteriori* p(θ| y^{obs}) de θ baseada nos dados observados y^{obs} – já que se determina uma distribuição *a priori* não-informativa, a distribuição *a posteriori* será um reflexo apenas daquilo que foi efetivamente observado (Enders, 2010);
2. Amostrar um valor θ* de p(θ| y^{obs});
3. Amostrar um valor y* de p(y^{mis}| y^{obs}, θ = θ*), a distribuição *a posteriori* condicional de y^{mis} dado que θ = θ*.

Os passos 2 e 3 são repetidos *m* vezes (Rubin, 1987; van Buuren *et al.*, 2006; White & Royston, 2009) e pode ocorrer uma pequena variação na geração dos bancos de acordo com o modelo de imputação escolhido.

Já tendo criado os bancos de dados imputados, a análise tradicional é realizada separadamente em cada um deles, gerando uma quantidade de estimativas pontuais tão grande quanto a quantidade de bancos de dados criados (*m*). A partir daí, a combinação dessas estimativas se dá a partir das chamadas "regras de Rubin". Assumindo Q como o parâmetro de interesse, ou seja, Q_i para *i* = 1,2, ..., *m* e que Q possa representar uma medida escalar – como média, correlação, coeficiente de regressão – ou um vetor de parâmetros, a combinação ($\bar{Q}_m$) das estimativas individuais ($\hat{Q}_i$) a partir da análise de cada banco de dados se dará pela equação (1.1). Para se determinar a variância combinada ($\bar{T}$ - equação 1.4) das estimativas considera-se a variância dentro das imputações ($\bar{U}_m$ - equação 1.2) e a variância entre elas (B_m - equação 1.3) (Rubin, 1987; van Buuren *et al.*, 1999).

$$\bar{Q}_m = \frac{\sum_{i=1}^{m} \hat{Q}_i}{m} \quad (1.1)$$

$$\bar{U}_m = \frac{\sum_{i=1}^{m} U_i}{m} \quad (1.2)$$

$$B_m = \frac{\sum_{i=1}^{m} \left(\hat{Q}_i - \bar{Q}_m\right)\left(\hat{Q}_i - \bar{Q}_m\right)}{\left(m-1\right)} \quad (1.3)$$

$$T_m = \bar{U}_m + \left(1 + m^{-1}\right) B_m \quad (1.4)$$

Além disto, para amostras razoavelmente grandes, o intervalo de 95% de confiança para $\bar{Q}_m$ será dado como na equação 1.5. As estimativas do risco relativo (ou do *hazard ratio*) podem ser determinadas com base na equação 1.6 e seus limites de confiança (95%) de acordo com a fórmula 1.7 (van Buuren *et al.*, 1999). Rubin (1987) também apresenta equações semelhantes, baseadas no teste *t*.

$$\bar{Q} \pm 1,96\sqrt{T_m} \quad (1.5)$$

$$\exp\left(\bar{Q}\right) \quad (1.6)$$

$$\exp\left(\bar{Q} \pm 1,96\sqrt{T_m}\right) \quad (1.7)$$

A utilização da IM requer a adoção de alguns pressupostos. Além de se assumir que o mecanismo de ocorrência dos dados faltantes é do tipo MAR – condição fundamental à realização de imputação múltipla, outra questão a ser considerada para a escolha do método de IM adequado é o padrão desta ocorrência – se monotônico ou não monotônico (Newgard & Haukoos, 2007; Nunes, 2007). Um pressuposto adicional mencionado por Harel & Zhou (2007) é a chamada congenialidade (*congeniality*) – caso o modelo de imputação e o de análise sejam os mesmos, os resultados obtidos com a imputação serão congeniais (*congenial*). Caso o modelo de imputação seja mais geral do que o de análise, os resultados

serão não-congeniais (*uncongenial*), mas ainda adequados – o que não ocorre se o modelo de análise for mais amplo do que aquele usado para imputar (Schafer, 1999).

Outro aspecto fundamental e complexo quando da realização de IM é a especificação do modelo de imputação, que abarca duas decisões: a forma do modelo (se linear, logístico, polinomial, etc.) e o conjunto de variáveis preditoras que o irão compor. Van Buuren e colaboradores (1999) afirmam que a função principal do modelo de imputação é gerar uma amplitude de valores plausíveis para aquelas variáveis sendo imputadas. A forma do modelo e seus parâmetros não são de grande interesse, o que torna a escolha exata da forma funcional do modelo de IM de menor importância.

A segunda questão a ser considerada diz respeito à seleção de co-variáveis do modelo de IM. O modelo deve ser capaz de preservar todas as associações importantes entre as variáveis (Patrician, 2002) e, deve incluir: as variáveis que farão parte do modelo de análise a ser aplicado posteriormente (incluindo o desfecho) (Moons *et al.,* 2006); variáveis auxiliares – que podem não fazer parte do modelo de análise, mas possuem alguma relação com os dados faltantes; e, variáveis que fazem parte do desenho do estudo (Haukoos & Newgard, 2007).

Alguns autores (van Buuren *et al.,* 1999; Spratt *et al.,* 2010; Camargos *et al.,* 2011) afirmam que uma estratégia de seleção inclusiva, que abarque a maior quantidade possível de variáveis, usando a maior quantidade possível de informação disponível, levará a um viés menor e precisão máxima. Outros autores (Allison, 2001; Patrician, 2002; Newgard & Haukoos, 2007) também mencionam que, na adição de co-variáveis, é melhor errar incluindo mais variáveis do que o necessário do que o contrário. Collins e colaboradores (2001), ao compararem estratégias inclusivas (com o maior número possível de variáveis) e restritivas (com o menor número possível), encontraram melhores resultados nos procedimentos de IM com abordagem inclusiva.

A ideia de se adicionar a maior quantidade de co-variáveis ao modelo tem relação com o fato de que, ao se incluir tantos preditores quanto possível, pode-se garantir mais facilmente de que o mecanismo dos dados faltantes seja MAR, reduzindo a necessidade de ajustes para considerar a possibilidade de MNAR (Schafer, 1997; van Buuren *et al.*,1999). Entretanto, em um banco de dados com centenas de variáveis, em um estudo de maior porte com diferentes blocos de questões que tratam de diferentes temas, como é o caso do Estudo Pró-Saúde, a adição de todas as variáveis disponíveis não é confiável e necessária - por conta de multicolinearidade e de problemas computacionais. Por mais que se incluam muitas co-variáveis, não haverá muitas mudanças nas estimativas obtidas após esse número passar de 15-25 (van Buuren *et al.,* 1999).

A literatura disponível sobre IM não fornece uma regra geral para guiar esta seleção de preditores, o que faz com que alguns autores optem por executar um procedimento *stepwise* para selecionar estas variáveis (Howard *et al.,* 2011). O que se afirma é que variáveis com distribuições altamente assimétricas podem fazer parte do modelo transformadas, para que se aproximem da normalidade (Patrician, 2002). As variáveis devem entrar de forma completa – isto é, categorizações devem ser feitas posteriormente, para que não haja restrição dos valores plausíveis a serem imputados (Newgard & Haukoos, 2007). O mesmo pode ser estendido para a imputação de escores – pode ser mais apropriado imputar as variáveis originais e depois reconstruir o escore com base nelas (Azur *et al.,* 2011). Além disto, a inclusão do desfecho de interesse do modelo de análise posterior parece reduzir o viés na imputação das suas co-variáveis (Moons *et al.,* 2006) e reter a associação entre o desfecho e estes preditores (que serão os desfechos do modelo de imputação) (He, 2010). Outro aspecto consensual é que variáveis candidatas a preditoras no modelo de IM não devem conter um alto percentual de dados faltantes (Nunes, 2007).

Uma abordagem para seleção de variáveis é sugerida por van Buuren e colaboradores (1999) e consiste em quatro passos:

1. Incluir todas as variáveis que aparecem no modelo de análise – se isto não for feito, haverá viés na análise posterior dos bancos imputados, principalmente se o modelo de análise contiver relações de predição fortes.
2. Incluir as variáveis relacionadas à (não) resposta – fatores que influenciam a ocorrência do dado faltante e outras variáveis de interesse cujas distribuições variem entre os grupos de respondentes e não-respondentes. Pode-se encontrar essas variáveis através do cálculo da correlação entre elas com uma indicadora (*dummy*) da não resposta naquela variável de interesse.
3. Adicionar variáveis que explicam uma porção considerável da variância da variável de interesse – preditores como esses ajudam a reduzir a incerteza das imputações. Também podem ser identificadas por meio da análise de correlação.
4. Por fim, remover dos conjuntos determinados nos passos 3 e 4 variáveis que tenham um alto percentual de dados faltantes dentro do subgrupo de casos incompletos.

Todos esses aspectos a serem considerados fazem com que a seleção das variáveis que irão compor o modelo de imputação se torne um passo importante, que deve ser cuidadosamente pensado e que considere não somente o modelo teórico-conceitual que subjaz o objeto de estudo, mas também o modelo capaz de predizer a ocorrência dos dados faltantes.

Quanto à decisão sobre a quantidade de banco de dados (*m*) a ser gerado, apesar de muitos autores (Kristman *et al.,* 2005; Harel & Zhou, 2007) utilizarem valores entre 3-10, Graham e colaboradores (2007) afirmam que o *m* necessário pode ser muito maior do que aquele que vem sendo proposto, e que se deve levar em conta o percentual de informação faltante – que difere do percentual de dados faltantes em situações mais complexas – e a eficiência que se deseja alcançar com a imputação, seguindo a equação proposta por Rubin (1987). Azur e colaboradores (2011) comentam que, na prática, gerar um *m* elevado pode ser problemático e demandar muito tempo, mas, também afirmam que a decisão sobre a quantidade deve se basear no tamanho do banco de dados, na quantidade de informação não observada e nos recursos computacionais disponíveis. Posteriormente, serão feitas considerações acerca da equação para determinação do *m* (seção 4.5.1).

A seguir, serão explicitadas as formas como os dados faltantes podem ser imputados, sob o título de "Modelos de Imputação Múltipla". Cabe ressaltar que não se tratam de modelos no sentido de forma funcional ou seleção de variáveis, conforme explicitado anteriormente. A seção diz respeito ao procedimento de imputação propriamente dito, que pode ocorrer de diferentes formas.

1.2.3.3 Modelos de Imputação Múltipla

He (2010) propõe uma classificação dos modelos de imputação múltipla em duas grandes categorias: *Joint Modelling* (JM) e *Sequential Regression Multiple Imputation* – ou, como mencionado por outros autores – *Fully Conditional Specification* (FCS) (van Buuren *et al.,* 2006).

A abordagem classificada como *Joint Modelling* (JM) divide as observações em grupos com características de não-resposta semelhantes e imputa os dados faltantes com cada padrão de acordo com um modelo conjunto para todas as variáveis que sejam comuns a todas as observações (He, 2010). Em outras palavras, JM envolve a especificação de uma distribuição multivariada para os dados faltantes e, na sequência, a amostragem de imputações da distribuição *a posteriori* condicional desses dados através de cadeias de Markov-Monte

Carlo (MCMC), por exemplo. Entre os exemplos clássicos estão o uso de modelos normais multivariados para variáveis contínuas, modelos log-lineares para variáveis categóricas e modelos de efeitos mistos para medidas repetidas ou análises multinível (Schafer, 1997). Sob uma distribuição *a priori* adequada, é possível derivar submodelos apropriados para cada padrão de dados faltantes, de onde as imputações serão amostradas. Van Buuren e Groothuis-Oudshoorn (2011) mencionam que o método de JM é atrativo caso uma distribuição multivariada forneça uma descrição razoável dos dados. He (2010) afirma que JM pode ser problemática por possuir pouca flexibilidade necessária para representar estruturas de dados complexas que são observadas em muitos estudos.

Dentre os métodos classificados como JM, Harel e Zhou (2007) trazem uma breve revisão de alguns deles: *Data Augmentation* (DA) – baseado em MCMC; uma variação do *hot deck* do método de imputação única; Aproximação bayesiana – ou *approximate Bayesian Bootstrap*; e, por fim, *Sampling importance/Resampling* (SIR).

Já entre os métodos de "*Fully Conditional Specification* (FCS)", o procedimento geral é caracterizado por modelos condicionais separados para cada variável incompleta, ou seja, o modelo de imputação é especificado separadamente para cada variável, utilizando as demais como preditoras. A cada passo do algoritmo da FCS, as imputações são geradas para os valores faltantes de uma variável. Esses valores imputados são usados na imputação da próxima variável, e o processo se repete até atingir a convergência (He, 2010). Um número pequeno de iterações geralmente é suficiente (van Buuren & Groothuis-Oudshoorn, 2011). FCS pode ser mais adequada do que os métodos de JM quando não há uma distribuição multivariada apropriada para lidar com o problema que se tem. Através da FCS pode-se facilmente acomodar características complexas dos dados em modelos de regressão construídos de acordo com os critérios aplicados nas análises de dados comuns, por exemplo: para variáveis contínuas, pode-se usar modelos lineares; para variáveis categóricas, modelos logísticos; e variáveis categóricas com mais de duas categorias, modelos multinomiais.

Van Buuren e colaboradores (2006) mencionam outros aspectos positivos e negativos da FCS. Comentam que a quantidade de iterações necessárias aos modelos de FCS pode ser menor do que aquela dos modelos de JM via MCMC e mencionam como principal vantagem do método de FCS a maior flexibilidade na hora de construir os modelos. E, apesar de se utilizar modelos de FCS (sobretudo MICE) em situações onde o padrão de dados faltantes é monotônico (Nunes *et al.*, 2009; Nunes *et al.*, 2010), o procedimento parece apresentar bons resultados quando o padrão é não-monotônico (van Buuren *et al.*, 2006). Um problema da

FCS é o fato de que, em algumas situações mais complexas, o modelo pode não convergir, tornando a aplicação do método inviável (Enders, 2010).

O procedimento de FCS já foi proposto sob diversos nomes, por diversos autores mencionados por van Buuren e Groothuis-Oudshoorn (2011): imputação através de regressão variável a variável; regressões sequenciais; pseudo-amostrador de Gibbs; MCMC parcialmente incompatível; imputação univariada iterativa; e, mais recentemente, equações em cadeia – *Multiple Imputation by Chained Equation* (MICE).

O procedimento de IM por equações em cadeia (MICE) vem sendo utilizado por diversos autores nos últimos anos (Nunes *et al.,* 2009; White & Royston, 2009; Nunes *et al.,* 2010; Azur *et al.,* 2011). Os primeiros relatos de aplicação da ideia datam de 1999, quando van Buuren e colaboradores (1999) afirmam o início do desenvolvimento de um procedimento que chamam de *regression switching*. Nos anos seguintes, métodos semelhantes e o próprio MICE foram desenvolvidos de maneira mais consistente. Da mesma forma que as demais técnicas de FCS, MICE constitui uma alternativa flexível para considerar diferentes distribuições de diferentes tipos de variáveis, que não podem ser agrupadas sob uma distribuição conjunta única, como nos métodos de JM.

Azur e colaboradores (2011) resumem o procedimento geral de imputação via equações em cadeia em quatro passos básicos. Os passos a seguir foram complementados pelo texto de White e colaboradores (2011):

> 1. Uma imputação única (como imputação por regressão ou pela média) é realizada em cada valor faltante no banco de dados. Esse valor imputado pode ser considerado como um "marcador de posição". Outra maneira de se "completar" inicialmente esses dados é através de uma amostragem aleatória simples dos valores observados, com reposição.
> 2. Os "marcadores de posição" de uma das variáveis – a primeira na sequência dos modelos determinados - (VAR 1) são removidos, e seus valores voltam a ser dados faltantes.
> 3. Os valores observados de "VAR 1" do passo 2 são regredidos nas outras variáveis do modelo de IM, que pode ou não consistir de todas as variáveis no conjunto de dados. "VAR 1" seria a variável dependente no modelo de regressão e todas as demais são as co-variáveis deste modelo. Os modelos de imputação operam sob os mesmos pressupostos dos modelos tradicionais – isto é, se for uma regressão linear, operará sob suas suposições. O mesmo se aplica para regressão logística, Poisson, etc.

> 4. Os valores faltantes de "VAR 1" são então substituídos pelas predições (imputações) do modelo de regressão – são substituídos por amostras da distribuição preditiva *a posteriori* desta variável. Esta amostragem pode se dar por meio do amostrador de Gibbs. Na sequência, quando "VAR 1" é utilizada como variável independente nos modelos de imputação de outras variáveis (VAR 2, VAR 3, ... , VAR *K*), tanto os dados observados quanto essas imputações são utilizadas.

Os passos 2-4 são repetidos para cada variável que tenha dado faltante. O ciclo de cada variável constitui uma iteração. No fim de um ciclo, todos os valores faltantes são substituídos com as predições oriundas das regressões que refletem as relações existentes na porção observada dos dados. As imputações são atualizadas e retidas ao fim de cada ciclo, o que gera um banco de dados imputado. A quantidade de ciclos pode ser determinada pelo pesquisador e, em geral, a convergência será alcançada quando os parâmetros que governam as imputações estejam estáveis (White *et al.*, 2011).

Uma distinção importante a se fazer diz respeito à forma como os valores imputados das variáveis substituem os dados faltantes – a substituição pode se dar via *Predictive Mean Matching* (PMM) ou via *Bayesian Linear Regression* (BLR). No método de PMM os parâmetros são estimados a partir de uma distribuição *a posteriori* própria e são calculados os valores preditos para os dados faltantes (y^{mis}) e para os observados (y^{obs}). Para completar os dados faltantes, procura-se a informação observada mais próxima de cada y^{mis} e se utiliza esse valor observado como o valor a ser imputado – em uma espécie de *hot deck*, selecionando como doadores valores semelhantes àqueles preditos pela amostragem da distribuição *a posteriori* (Rubin, 1987; Nunes *et al.*, 2009). Já no BLR, os valores imputados são as próprias predições de y^{mis}. Nos programas estatísticos, pode-se determinar se o procedimento a ser adotado é o PMM ou o BLR (van Buuren & Groothuis-Oudshoorn, 2011), e, tal decisão é importante sobretudo se a variável a ser imputada é contínua.

A característica mais importante do MICE é sua habilidade em lidar com diferentes tipos de variáveis, cada uma com seu próprio modelo de imputação, que é atualizado com base no modelo anterior, e utilizado para determinar o modelo posterior. Zhou e colaboradores (2001) também afirmam que o procedimento evita que ocorra extrapolação dos limites (amplitude) dos dados e, o fato de se usar amostras da distribuição preditiva *a posteriori* ajuda a preservar distribuições e associações entre as variáveis. A maior complexidade do procedimento reside na necessidade de se especificar modelos de imputação

distintos para cada variável, diferentemente da abordagem de JM, na qual um único modelo de imputação é especificado (He, 2010; Azur *et al.*, 2011).

1.2.3.4 Imputação Múltipla em estudos de sobrevivência

Na Epidemiologia, poucos são os trabalhos dedicados a lidar com a questão da IM em estudos de sobrevivência de maneira mais metodológica, independente do modelo de análise utilizado (paramétrico, semi-paramétrico ou não paramétrico). No que tange à seleção das variáveis a compor o modelo e ao procedimento geral de imputação, a técnica ocorre de maneira semelhante às demais e parece fornecer resultados satisfatórios ao se aplicar MICE (van Buuren *et al.*, 1999; White & Royston, 2009).

A particularidade quando da construção dos modelos de imputação em sobrevivência diz respeito à inclusão do desfecho: nos modelos de regressão, deve-se considerar o status (censura ou caso) e o tempo de contribuição de cada participante ao estudo como desfechos, e não só o status do indivíduo.

Alguns autores (van Buuren *et al.*, 1999; White & Royston, 2009) discutem de que maneira o tempo deve fazer parte dos modelos de imputação, já assumindo como necessária a adição do desfecho nos modelos, a fim de evitar diluição da associação entre as co-variáveis e o desfecho do modelo de análise (neste caso, a sobrevivência), conforme relatam Moons e colaboradores (2006). Afirmam que essa inclusão pode se dar de diversas formas: adicionando o status, o tempo e o logaritmo do tempo como preditores no modelo de imputação; adicionando apenas o status e o logaritmo do tempo; e outras. Em outro trabalho recentemente publicado (Ali *et al.*, 2011), os autores comparam modelos de IM que incluem o tempo de sobrevivência como preditor com modelos que não o incluem e mencionam que, semelhante ao descrito na literatura, a exclusão do desfecho nos modelos de IM leva a viés em direção ao nulo nos coeficientes do modelo de regressão e melhores resultados foram obtidos nos modelos de IM que consideravam o tempo de sobrevivência.

Ainda não há um consenso sobre a forma desta inclusão (White & Royston, 2009) e há autores que sequer mencionam de que forma o desfecho foi incorporado aos modelos de imputação (Baneshi & Talei, 2011). Em análises que empregam o modelo de riscos proporcionais de Cox, há indícios de que o risco basal ($H_0(t)$) deva ser estimado e incluído no modelo (Paik, 1997). White & Royston (2009) consideram três possíveis métodos de

estimação de $H_0(t)$: conhecimento prévio – utilizando um valor de um estudo anterior, por exemplo – abordagem que não é aconselhável; pelo método de Nelson-Aalen; e pelo método de Cox, estimando $H_0(t)$ iterativamente. Os autores comparam modelos com $H_0(t)$ estimado por esses métodos e concluem que o estimador de Nelson-Aalen é o que apresenta melhores resultados como primeira escolha, mas o modelo de imputação deve ser cuidadosamente pensado, de forma a incluir a melhor estimativa de $H_0(t)$, de acordo com o cenário que se tem.

Desta forma, além das considerações quanto aos pressupostos necessários à adoção de IM, à forma funcional do modelo e as variáveis que o irão compor, à quantidade de bancos de dados a ser gerado (*m*), e, quanto à maneira que a imputação ocorrerá – se via JM ou FCS, a forma como o tempo de sobrevida será incorporado ao modelo constitui a principal distinção da IM em modelos de sobrevivência, exigindo atenção especial do analista.

2 **JUSTIFICATIVA**

O presente trabalho está sendo proposto para solucionar um problema identificado no Estudo Pró-Saúde, relacionado à ocorrência de dados faltantes, sobretudo em variáveis relacionadas a características sócio-econômicas e a questões relacionadas à Epidemiologia ao longo do curso de vida. Como mencionado anteriormente, apesar dos esforços para se evitar a não-resposta durante uma mesma fase e perdas durante o seguimento da coorte, o EPS apresenta percentuais de dados faltantes expressivos em determinadas variáveis e, acredita-se que a imputação múltipla seja uma ferramenta capaz de ajudar a lidar com essa questão.

Infelizmente, o material disponível sobre o assunto ainda está concentrado na forma de textos metodológicos que requerem um conhecimento profundo da teoria estatística, e, por vezes, de Inferência Bayesiana (Rubin, 1987). Mais ainda, os trabalhos existentes no Brasil ainda são poucos, e a aplicação em determinadas áreas da Epidemiologia ainda é incipiente. Muito provavelmente, a pouca aplicação se deve ao fato da dificuldade imposta quando da escolha e utilização da técnica por epidemiologistas com conhecimentos limitados de Estatística.

Facilitar o entendimento do assunto por pesquisadores da área de saúde constitui o principal desafio aqui proposto. Criar um texto metodológico básico, que "traduza" os princípios fundamentais da imputação de dados e de que maneira ela pode ser aplicada em nosso meio é uma tarefa importante para garantir que se use a técnica com maior convicção e segurança. Além disto, o uso de um *software* livre para execução desta tarefa é outro aspecto que garante o acesso de todos ao material a ser disponibilizado posteriormente.

Um outro ponto importante é o fato de pouco se ter explorado o tema da imputação múltipla em estudos que envolvam técnicas de análise de sobrevivência, bastante adequadas para se tratar diversos problemas em Epidemiologia (Carvalho *et al.,* 2011). Elaborar um texto na forma de tutorial para aplicar IM em modelos de sobrevivência atende não só uma necessidade do EPS, como também preenche uma lacuna do conhecimento identificada na literatura brasileira e internacional.

Em relação à última, um aspecto a ser considerado é o fato de que, mesmo em estudos internacionais publicados em periódicos importantes, alguns problemas na utilização do método de IM podem ser apontados, a saber: a não determinação do número de bancos de dados imputados a serem gerados – muitos estudos determinam o valor de "*m*" sem um embasamento claro (Kristman *et al.,* 2005; Harel & Zhou, 2007; Howard *et al.,* 2011), desconsiderando a existência de uma fórmula que pode ser empregada no seu cálculo e que,

segundo esta fórmula, para um elevado percentual de dados faltantes, um "*m*" maior será necessário para que a eficiência relativa a se alcançar com a IM seja adequada (Rubin, 1987; Graham *et al.,* 2007); o uso de aplicativos pagos na execução das técnicas e das análises (Sinharay *et al.,* 2001; Buhi *et al.,* 2008) ou ainda o uso de pacotes ou programas antigos, que não correspondem a prática dos dias atuais (Sinharay *et al.,* 2001); a aplicação de técnicas de IM em dados faltantes que não podem ser assumidos como MAR (ou seja, que são sabidamente MNAR), o que inviabilizaria a correta utilização do método (Mishra & Dobson, 2004; Kristman *et al.,* 2005); e, por fim, alguns estudos não aplicam a IM a um cenário de dados reais – ocorre simulação dos dados e da não-resposta e não o uso de um banco efetivamente coletado, que demonstre relações reais entre as variáveis sob análise (Buhi *et al.,* 2008).

Tendo em vista estas lacunas e a necessidade levantada por outros autores brasileiros (Nunes *et al.,* 2010), que mencionam a necessidade de material metodológico nesta área, este trabalho visa fornecer subsídios e colaborar para divulgação e uso eficiente da imputação múltipla de dados faltantes em pesquisas epidemiológicas.

3 **OBJETIVOS**

- Rever análises realizadas no âmbito do Estudo Pró-Saúde, passíveis de simulação, a fim de se verificar cenários mais apropriados para a aplicação da IM (artigo 1);
- Desenvolver um tutorial para imputação múltipla de dados em estudos epidemiológicos (artigo 2).

4 **MATERIAIS E MÉTODOS**

Para criação de um protocolo para imputação múltipla, optou-se por utilizar uma análise recentemente concebida do Estudo Pró-Saúde, na qual se tenha identificado um percentual de dados faltantes inferior a 5% dentro dos modelos de análise utilizados, isto é, a combinação do percentual de não-resposta de cada variável deveria resultar em um percentual de dados faltantes total menor que 5% entre as variáveis explicativas e, com isso, após a remoção dos indivíduos com estas informações não observadas, o banco seria considerado "completo". A adoção de 5% como critério de decisão foi arbitrária e necessária para que se decidisse por um modelo dentre os diversos disponíveis. Neste exemplo selecionado, a técnica de análise de sobrevivência - método estatístico bastante utilizado em Epidemiologia - deveria ter sido aplicada.

A partir do banco de dados utilizado nesta análise de referência, foram conduzidas simulações de não-resposta, que, ao serem imputadas, permitiram comparar a eficiência da imputação múltipla em relação à análise de observações completas, tendo como "padrão-ouro" o banco de dados original, considerado completo.

A seguir, são descritas as bases metodológicas empregadas, as características do estudo, bem como a análise revisitada e as variáveis utilizadas na mesma. Logo depois, são apresentados os métodos empregados na construção do banco de dados e simulações/imputações deste estudo. Cabe ressaltar que o propósito deste trabalho não é debater sobre o modelo teórico-conceitual utilizado ou acerca de associações encontradas ou não. Trata-se de um exemplo de aplicação com fins ilustrativos da técnica de imputação múltipla neste cenário.

4.1 **Bases metodológicas**

4.1.1 O modelo de sobrevivência

Modelos de sobrevivência podem ser definidos como "uma classe de modelos quantitativos estocásticos utilizada para analisar características e fatores associados ao tempo até a ocorrência do desfecho de interesse" (Carvalho *et al.,* 2011). Tratam-se dos modelos de escolha quando é possível incorporar a informação do tempo na análise de dados. Como

quaisquer outros modelos de regressão, são compostos por uma variável resposta (desfecho), variáveis explicativas, uma função de ligação e a estrutura de erro.

Através de tais modelos, pode-se responder a questões como a probabilidade de um indivíduo sobreviver por mais de "t" unidades de tempo; o risco de sofrer um evento no tempo "t", dado que o indivíduo sobreviveu até aquele momento; o risco de sofrer um evento até um determinado tempo "t", dentre outras.

Os modelos de sobrevivência podem ser divididos em três grandes grupos: modelos não paramétricos – e, entre eles, os mais utilizados são o estimador produto de Kaplan-Meier e o estimador de Nelson-Aalen e os testes de log-rank e Peto; modelos paramétricos – utilizando distribuições exponenciais, Weibull ou lognormal; e modelos semi-paramétricos – sendo o modelo de riscos proporcionais de Cox (e suas extensões) o mais utilizado na área da saúde (Hosmer *et al.*, 2011).

Este último modelo – o de riscos proporcionais de Cox – foi utilizado neste trabalho quando da ilustração das técnicas de IM. O modelo, proposto na década de 1970 (Cox, 1972), assume que as co-variáveis têm um efeito multiplicativo na função de risco e a razão entre o risco de ocorrência do evento para dois indivíduos é constante ao longo no tempo – o que explica o termo "riscos proporcionais". Através deste pressuposto de proporcionalidade, é possível estimar os efeitos das variáveis explicativas, sem que seja necessário supor qualquer distribuição para o tempo de sobrevivência (Hosmer *et al.*, 2011). É considerado semi-paramétrico por não supor qualquer distribuição para a função de risco basal e sua porção paramétrica diz respeito ao pressuposto de que as variáveis explicativas exercem efeito multiplicativo sobre o risco. No modelo de Cox, a estimação dos coeficientes do modelo se dá através de verossimilhança parcial e a qualidade do ajuste pode ser verificada de diversas maneiras – através dos resíduos de Schoenfeld, martingale, escore, deviance, etc.

Para aplicação correta da análise de sobrevivência, independente do modelo a ser escolhido, duas questões principais devem ser observadas: detalhar o máximo possível a informação relacionada ao tempo (incluindo sempre datas de ocorrência dos fenômenos de interesse) e montagem cuidadosa do banco de dados (Carvalho *et al.*, 2011). Tais aspectos serão mencionados quando da análise dos dados. Para que se tenha informação correta quanto ao tempo de ocorrência dos fenômenos, necessário à análise, o modelo de sobrevivência requer um acompanhamento dos indivíduos, seja através de um estudo de seguimento ou, de maneira menos precisa, em um estudo seccional, ao se obter alguma forma de se resgatar o histórico dos participantes, com base em informações em prontuários médicos, por exemplo.

Neste trabalho, a informação sobre o tempo de seguimento de cada indivíduo foi reconstruída com base nos relatos dos participantes, de acordo com critérios pré-estabelecidos. O fato de não se ter acompanhado efetivamente os indivíduos pode simbolizar um acréscimo de incerteza ao estudo, porém, não o invalida. Novamente, não cabe levantar considerações sobre o objeto de estudo em si, mas sobre a eficiência da IM em lidar com o problema dos dados faltantes na análise de sobrevivência.

4.2 **O Estudo Pró-Saúde – características gerais**

O Estudo Pró-Saúde (EPS) é um estudo longitudinal realizado entre funcionários técnico-administrativos de uma universidade localizada no Rio de Janeiro, cujo início se deu em 1998. Entre seus objetivos principais figuram a investigação do papel de determinantes sociais no estado de saúde dos indivíduos (Faerstein *et al.*, 2005).

Encontra-se atualmente em sua quarta etapa, sendo as anteriores realizadas nos anos de 1999 (Fase 1), na qual participaram 4020 trabalhadores (91% dos elegíveis); 2001 (Fase 2), na qual a taxa de participação foi de 83% (3.574 dos 4.317 elegíveis) (Faerstein *et al.*, 2005). As fases 1 e 2 são as chamadas "linhas de base", e 3.253 funcionários participaram de ambas. Em 2006, ocorreu a Fase 3 do estudo, na qual participaram 3604 indivíduos, e, destes, 3058 participaram da linha de base, representando uma cobertura de 94% daqueles inicialmente acompanhados. A Fase 4 teve início em setembro de 2011 e tem previsão de ser concluída em 2013. Para a realização deste trabalho, foram usadas as informações provenientes da linha de base (Fases 1 e 2) somente.

Dentre todos os funcionários técnico-administrativos da universidade convidados a participar do estudo, eram elegíveis apenas aqueles que não estivessem cedidos a outras instituições ou licenciados por motivos não relacionados à saúde. Detalhes sobre a população de estudo encontram-se em Faerstein e colaboradores (2005).

Uma característica importante é a maneira como se busca evitar perdas neste estudo, um aspecto comum em estudos de seguimento de qualquer porte. Faerstein e colaboradores (2005) mencionam cinco principais estratégias para retenção e manutenção de contato com os participantes, sobretudo aqueles da linha de base, a saber: seguimento de indivíduos com maior potencial de retenção; comunicação e divulgação, contando inclusive com recursos da própria universidade; inclusão, sempre que possível, de interesses e expectativas da comunidade; atualização regular de informações cadastrais, através do apoio de setores

administrativos; seguimento remoto através do relacionamento com bancos de dados secundários.

Mesmo com todas essas estratégias, ainda se pode observar a ocorrência de não-resposta entre os dados do EPS. Isto decorre principalmente do fato de os questionários serem autopreenchíveis, e, muitas vezes, o participante esquecer de responder alguma pergunta ou a "pular" acidentalmente. Há ainda situações nas quais o participante ignora ou prefere não fornecer as informações requisitadas no questionário. Outra razão que leva à perda dos dados é a marcação de mais de uma opção de resposta, impossibilitando a identificação de qual delas é a adequada, e, consequentemente, levando ao descarte da informação daquele indivíduo naquela variável. Além disto, a perda entre as fases da coorte também ocorre, já que alguns funcionários se recusam a participar novamente do estudo por diversas justificativas, porém, a equipe do EPS sempre busca diversas maneiras de reverter tais recusas.

4.2.1 Coleta de dados e qualidade da informação

A coleta de dados do EPS ocorre com o auxílio de pesquisadores treinados e supervisores, no local de trabalho do indivíduo. Por meio de questionários autopreenchíveis, são verificadas questões quanto a condições sócio-econômicas, cor/raça, mobilidade geográfica e social, experiência de discriminação, estresse no trabalho, padrões de rede e apoio social, aspectos de saúde da mulher, morbidades, acidentes do trabalho, transtornos mentais comuns e de comportamentos relacionados à saúde (atividade física, padrões de dieta e tabagismo, utilização de procedimentos, serviços e medicamentos), e, são aferidos peso, estatura, circunferência abdominal e pressão arterial, por equipe treinada, utilizando métodos padronizados e com controle de qualidade regular (Faerstein *et al.*, 2005).

Nas duas etapas da linha de base foram realizados pré-testes a fim de se avaliar a estrutura e adequação do questionário (clareza, constrangimento em algumas questões, sequência, transição entre os blocos de perguntas). Também se testou aspectos referentes ao processamento de dados e, sempre que possível, a avaliação ocorria entre voluntários semelhantes àqueles da população fonte. Além disto, ocorreram estudos-piloto entre os funcionários que não pertenciam ao quadro efetivo da unidade, a fim de não apenas testar as etapas de coleta dos dados na linha de base, mas também a confiabilidade dos instrumentos. A repetição da aplicação dos questionários ocorria em um intervalo de duas semanas (Faerstein

et al., 2005; Boclin, 2011). O EPS foi aprovado pelo Comitê de Ética em Pesquisa da instituição na qual é conduzido.

A seguir, as variáveis utilizadas e o procedimento de coleta das mesmas no exemplo escolhido serão descritos.

4.3 **Estudo de referência**

Selecionou-se a análise realizada por Boclin (2011), em tese de doutorado recentemente defendida. Uma das propostas de tal trabalho foi averiguar se posição sócio-econômica durante a infância, início da vida adulta, ou ao longo da vida era mediadora da relação entre cor/raça e a ocorrência de miomas uterinos (MU) na população.

Para conduzir tal investigação, utilizou-se os dados da linha de base do EPS para a população feminina, que contava com 1819 participantes, dentre 2466 funcionárias elegíveis (73,8% das elegíveis foram avaliadas em 1999-2001). No trabalho selecionado, incluíram-se apenas as mulheres que continham todas as informações para as variáveis consideradas no modelo. Com isto, o percentual de dados faltantes atingiu 18,6% (exclusão de mulheres sem informação sobre alguma variável de exposição e daquelas sem informação sobre o desfecho ou tempo de ocorrência do mesmo).

4.3.1 Variáveis utilizadas

4.3.1.1 Desfecho analisado

O desfecho utilizado na análise foi o diagnóstico médico auto-relatado de miomas uterinos – avaliado por meio da questão: "*Alguma vez um médico lhe informou que você tinha mioma uterino, um tumor benigno no útero?*", cujas respostas possíveis eram "Sim" e "Não", tratando-se, portanto, de variável dicotômica.

Outras informações coletadas acerca dos miomas uterinos foram a idade da participante quando do diagnóstico do mioma e sobre a realização de cirurgia de retirada do útero (histerectomia) e a idade da mulher quando submetida a esta, para que se reconstituísse adequadamente o período de seguimento das participantes.

4.3.1.2 Variável de exposição (principal)

A cor/raça das participantes, utilizada como variável de exposição principal, foi coletada através da pergunta aberta "*Em sua opinião, qual é a sua cor ou raça?*". Após agrupamento dos relatos das participantes, a variável em questão foi categorizada da seguinte maneira: branca, parda, preta e amarela. A categoria amarela foi excluída das análises do estudo de base por apresentar um número pequeno de participantes (n=8, 0,44%). Para este trabalho, também se optou por excluir mulheres de cor/raça amarela, para que se pudesse comparar os resultados com aqueles do estudo de base e por simbolizarem um grupo pequeno de indivíduos, que poderiam introduzir 'ruídos' às análises.

4.3.1.4 Outras variáveis utilizadas

Outras variáveis utilizadas no modelo como marcadoras de acesso e utilização de serviços de saúde foram: Plano de saúde; Realização de teste Papanicolaou; Realização de exame de mama.

A variável "plano de saúde" foi coletada em três categorias: 1 - Sim, como titular; 2 - Sim, como dependente; 3 – Não, e, posteriormente, dicotomizada. As variáveis "realização de teste Papanicolaou" e "realização de exame de mama" foram coletadas em quatro categorias: 1 – Nunca fiz o exame; 2 – Há mais de 3 anos; 3 – Entre 1 e 3 anos atrás; 4 – há menos de 1 ano. Estas foram dicotomizadas como 'realizou há menos de 3 anos' e 'nunca realizou ou realizou há mais de 3 anos'.

A escolaridade da participante também fazia parte dos modelos em questão, tendo sido coletada em sete categorias e, posteriormente, transformada em três: 0 - até 1º grau completo; 1 – até 2º grau completo e 2 – universitário completo ou mais.

Outra variável importante foi a idade da participante, que compunha os modelos de análise do estudo de referência de forma contínua e, para a simulação, imputação e análise deste trabalho foi inserida na forma de tercis. A decisão pelos tercis se deu principalmente por questões práticas na simulação dos dados faltantes.

Ainda na simulação (a ser detalhada na seção 4.5.1), a variável referente à cor/raça de acordo com a classificação proposta pelo IBGE, coletada através da pergunta *"O Censo Brasileiro (IBGE) usa os termos preta, parda, branca, amarela e indígena para classificar a cor ou raça das pessoas. Se você tivesse que responder ao Censo do IBGE hoje, como se*

classificaria a respeito de sua cor ou raça?”, cujas respostas possíveis eram: preta, parda, branca, amarela e indígena, também foi utilizada.

4.3.2 Análise estatística adotada

Para lidar com o problema de os dados terem sido coletados transversalmente (dados da linha de base do EPS), buscou-se resgatar as histórias de seguimento das participantes, a partir das informações por elas relatadas.

Assumiu-se que o seguimento se iniciou aos 20 anos de idade para todas as mulheres e o tempo final de “acompanhamento” foi determinado da seguinte maneira: para mulheres que não apresentavam miomas uterinos, utilizou-se a idade em 1999; para as mulheres que relataram miomas, a idade do diagnóstico do mesmo, e, para mulheres que relataram histerectomia, a idade da cirurgia. Determinou-se como tempo final de seguimento máximo a idade de 50 anos. Desta forma, dez casos de miomas diagnosticados após esse período foram censurados.

Nas análises multivariadas, foram utilizados modelos de riscos proporcionais de Cox para estimar a Razão de Hazards com intervalos de 95% de confiança (IC 95%) (Boclin, 2011). O modelo utilizado neste trabalho é o que tem por desfecho o status da participante (caso - 1; censura – 0) e o tempo de seguimento conforme os critérios antes mencionados. Como variáveis de exposição estão a cor/raça e a idade das participantes e como fatores de confusão do modelo, as variáveis marcadoras de acesso aos serviços de saúde (plano de saúde, exame de Papanicolaou, exame de mama) e a escolaridade da mulher.

4.4 **Organização do banco de dados**

Utilizando o mesmo conjunto de dados utilizado por Boclin (2011), procedeu-se a exclusão das participantes consideradas inelegíveis para este estudo: aquelas com informação faltante no desfecho sob investigação – por se tratar de um modelo de sobrevivência, foram excluídas mulheres sem informação sobre a ocorrência de mioma uterino ou sem informação sobre a idade do diagnóstico do mesmo. Além disto, optou-se por excluir as mulheres que afirmaram ter realizado histerectomia, mas não informaram a idade da mesma e aquelas que não informaram sobre realização ou não desta cirurgia, já que não se podia determinar a

extensão do seu período sob risco. Em seguida, foram removidas as mulheres cujo mioma uterino ou a histerectomia ocorreram antes dos 20 anos de idade, já que haviam sido consideradas inelegíveis no estudo de base. Por fim, mulheres de cor/raça amarela foram excluídas.

Com as exclusões acima mencionadas, o banco de dados inicial, com 1819 mulheres, passou a ser constituído de 1593 participantes. Destas, 5,5 % (n=88) não tinham informação em alguma das variáveis do modelo proposto ou outras variáveis a serem utilizadas durante os procedimentos de simulação ou imputação. Com a remoção destas últimas, o banco de dados considerado completo (submetido a simulação nesta dissertação) consistiu de 1505 mulheres, todas com informações completas no desfecho, idade, cor/raça, plano de saúde, realização de exame de mama e Papanicolaou, escolaridade e cor/raça de acordo com a classificação proposta pelo IBGE (Figura 2).

Este banco de dados foi organizado de forma a permitir uma análise utilizando o modelo de Cox, e, para isso, foram acrescentadas colunas referentes ao status do indivíduo (1- se caso; 0 – se censura), quanto ao tempo inicial (20 anos) e final de seguimento de cada participante, e, quanto ao tempo de seguimento de cada mulher. Toda a montagem do banco de dados foi feita utilizando o *software* R, versão 2.15 (R Development Core Team, 2012). Os scripts utilizados se encontram no apêndice 8.2.1 (p.120).

4.5 **Estudo de simulação**

4.5.1 Simulação dos dados faltantes

Para este estudo, o percentual de dados faltantes do modelo utilizado (aproximadamente 5% após a combinação das variáveis explicativas, conforme decisão anteriormente mencionada) permitiu a obtenção de um banco de dados com informação completa nas co-variáveis de interesse e no desfecho. Este banco foi assumido como referência e analisado através do modelo de riscos proporcionais de Cox. Os resultados desta análise são considerados “padrão-ouro” quando da comparação com os resultados obtidos na IM. A descrição geral deste conjunto de dados (análise univariada) pode ser encontrada no apêndice 8.1 (p. 118).

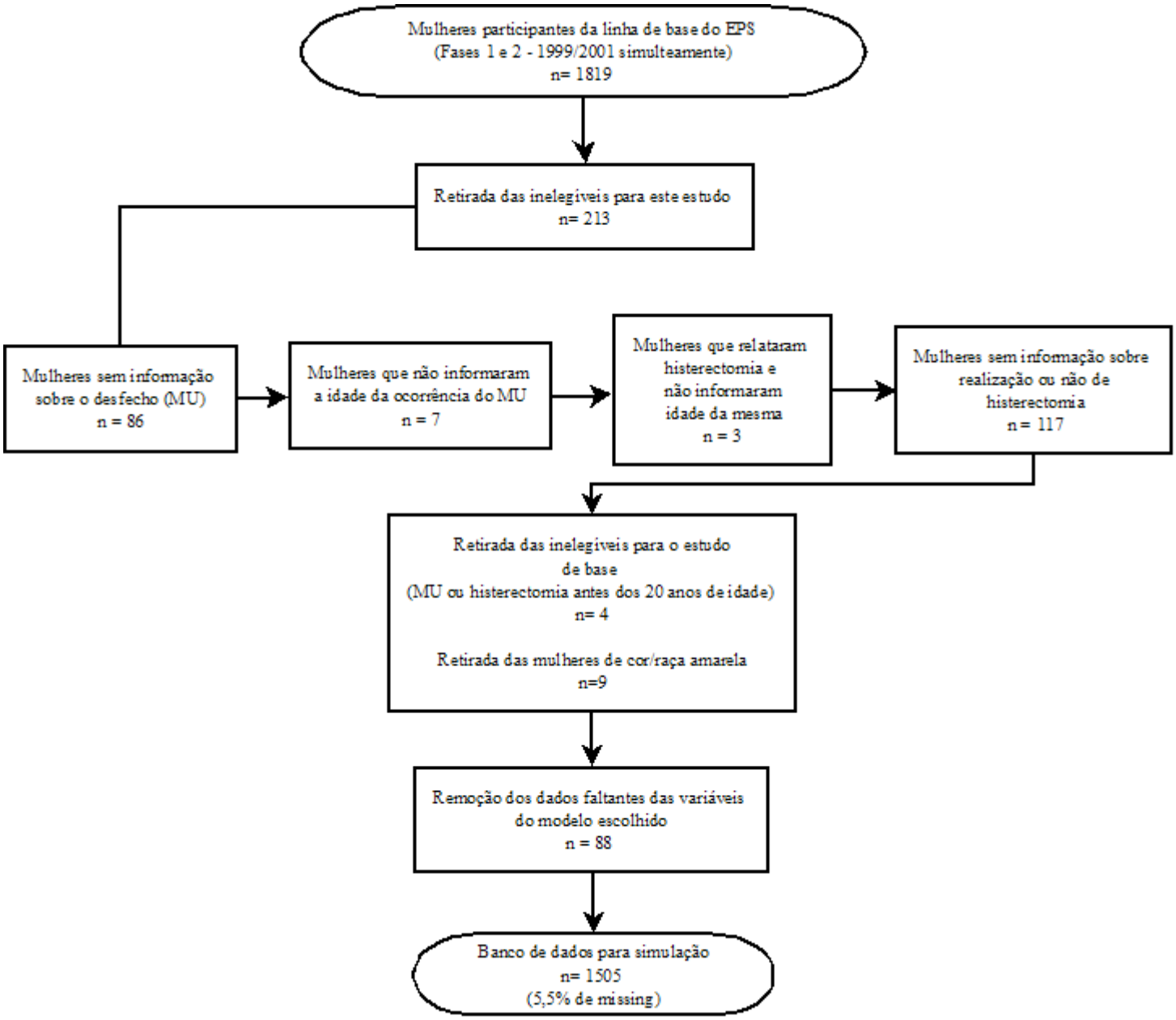

Figura 2 – Fluxograma referente à montagem do banco de dados no qual se simulou a ocorrência de dados faltantes e posterior imputação múltipla – dados do Estudo Pró-Saúde – RJ (1999-2001).

Na sequência (Figura 3) foi simulada não-resposta nestes dados. Fixou-se o padrão de dados faltantes em não-monotônico, já que esse era o padrão observado entre os dados antes da remoção dos 5,5% faltantes. A não-resposta foi simulada na variável cor/raça das participantes (variável multinomial), já que, além de constituir a exposição principal do estudo de base, tem se observado, atualmente, na quarta fase no EPS, um aumento no percentual de dados faltantes nesta variável (dados não publicados). Com isso, o tutorial gerado terá função imediata quando da análise desses dados provenientes da Fase 4 do EPS.

Ocorreu variação no percentual de dados faltantes – de um cenário com menor percentual (5%) a valores mais elevados (10, 20 e 30%); no mecanismo de ocorrência dos dados faltantes – MCAR, MAR e MNAR; e, quando da imputação, houve variação na quantidade de banco de dados (*m*) a ser criado com o procedimento - m= 5, 10, 20 e 100. Considerou-se suficiente m=100, já que para alcançar uma eficiência relativa de 100%, seria necessário $m \to \infty$, de acordo com a equação proposta por Rubin (1987) (equação 4.1) e um *m* elevado levaria a uma situação computacionalmente intensiva. Mesmo com um percentual de dados faltantes elevado, m=100 é capaz de fazer com que a eficiência relativa a ser obtida com a imputação seja satisfatória.

$$Eficiência = \left(1 + \frac{\gamma}{m}\right)^{-\frac{1}{2}} \quad (4.1)$$

Para determinar quais indivíduos deveriam ser tansformados em dados faltantes em cada cenário, determinou-se um modelo logístico para predição da probabilidade da ocorrência de dados faltantes na variável cor/raça (ou seja, desfecho: 1 – ser dado faltante para cor/raça; 0 – não o ser), utilizando o banco de dados com n=1593 participantes, que possuía 3,2% de dados faltantes na mesma. Um procedimento semelhante é descrito por Zhou e colaboradores (2001).

O modelo logístico no cenário MAR contou com as variáveis idade em tercil e escolaridade das participantes (em três categorias), já que pareciam capazes de explicar, em parte, a ocorrência do desfecho em questão (ou seja, ser ou não dado faltante para cor/raça). Com isto, havia 9 padrões possíveis de co-variáveis, e, dentro de cada um deles, determinou-se a probabilidade de ocorrência de dados faltantes para cor/raça. Estas probabilidades associadas a cada padrão foram então aplicadas aos mesmos padrões no banco de dados completo (n=1505) e, multiplicando-se estes valores pela quantidade de participantes em cada

padrão, pôde-se determinar quantos indivíduos deveriam ser removidos em cada um deles, a fim de se gerar 5, 10, 20 e 30% de dados faltantes na totalidade.

Para o cenário MNAR, além de idade e escolaridade, incluiu-se a variável cor/raça de acordo com a classificação do IBGE, a fim de se forçar a associação entre o desfecho (ser ou não dado faltante) e a (não) resposta à questão sobre sua cor/raça. Isto levou à existência de 18 padrões de co-variáveis, que tiveram suas probabilidades de ocorrência do desfecho calculadas de forma semelhante ao cenário anterior, e, novamente, foram multiplicadas pela quantidade de indivíduos em cada padrão e recalculadas para gerar 5, 10, 20 e 30% de não-resposta.

Para MCAR, foi realizada uma amostragem aleatória utilizando os números de identificação das participantes, obedecendo as quantidades necessárias para gerar 5% (n=76), 10% (n=152), 20% (n=301) e 30% (n=452) de não-resposta. Os indivíduos sorteados foram então transformados em dados faltantes na variável cor/raça.

A partir de então, já se conhecendo quantos e quais indivíduos deveriam ter a informação sobre sua cor/raça 'apagada' do banco de dados completo, procedeu-se a simulação seguida da imputação de dados, a ser descrita na próxima seção.

4.5.2 Imputação Múltipla

Os bancos de dados gerados com as simulações foram imputados pela técnica de IM por equações em cadeia (MICE). O modelo de imputação foi construído conforme a seleção de variáveis proposta por van Buuren e colaboradores (1999), considerando as particularidades de cada variável disponível. Foram incluídas na matriz de predição da cor/raça das participantes as variáveis: idade (em tercis), escolaridade (três categorias), plano de saúde, realização de exame de mama, realização de exame de Papanicolaou (dicotomizadas), status da participante (se caso ou censura) e tempo de contribuição no estudo (em anos).

Apesar da recomendação de se inserir variáveis de forma completa no modelo de imputação (e realizar recategorizações posteriormente), optou-se por adicionar algumas das variáveis necessárias já recategorizadas, a fim de se evitar problemas de convergência na imputação ou na análise sequencialmente conduzida.

O modelo de análise já havia sido estabelecido *a priori* no estudo de base, e foi repetido em cada banco de dados imputado, utilizando o modelo de riscos proporcionais de

Cox da mesma maneira: como desfecho utilizou-se o status e o tempo; como variáveis de exposição a cor/raça das participantes já imputada e como co-variaveis a idade – em tercis, plano de saúde, realização de exame de mama e Papanicolaou (dicotomizadas) e a escolaridade da participante (em 3 categorias). Para evitar problemas de convergência neste modelo, o número máximo de iterações foi alterado para 100 (o R possui como padrão 20 iterações). Ao final das análises, os valores de cada banco de dados foram combinados, obedecendo as "regras de Rubin" anteriormente expostas.

Para determinar a distribuição das estimativas obtidas em cada cenário de imputação, o processo foi replicado 100 vezes (N – figura 3) e, ao final, foi determinada a média Monte Carlo dos coeficientes do modelo de regressão para a variável de interesse (cor/raça) e seus erros-padrão, bem como a variância entre os coeficientes.

Utilizando as médias dos coeficientes e erros, determinou-se a Razão de Hazards e seu intervalo de confiança (IC95%) em cada cenário, apenas para a variável de exposição principal, a fim de compará-los com os valores do "padrão-ouro" e com os da análise de observações completas. Estas últimas também foram replicadas 100 vezes, a fim de se determinar a distribuição das estimativas desta análise, após simulação dos dados faltantes, da mesma forma como ocorreu com a IM.

A análise foi conduzida no programa estatístico R (R Development Core Team, 2012), versão 2.15. Ao final (apêndice 8.2, p.120) são apresentados os scripts utilizados neste trabalho. A decisão pelo R se deu principalmente por se tratar de um *software* livre, de fácil acesso por qualquer pesquisador, e que cobre de maneira satisfatória as análises necessárias tanto na simulação e na IM quanto na análise de sobrevivência.

O principal pacote do R para realizar imputações por equações em cadeia é o "MICE", criado por van Buuren e Groothuis-Oudshoorn (2011). O "MICE" permite a aplicação da técnica em diferentes cenários, com diferentes tipos de variáveis (normais, binárias, multinomiais, etc), bem como a combinação das estimativas geradas (pelas "regras de Rubin"), diagnóstico dos modelos de imputação, alterações na matriz de predição a ser utilizada para IM, além da determinação de quais variáveis do banco de dados serão imputadas e de que maneira o serão. Também inclui a possibilidade de análise de sensibilidade para se considerar um mecanismo de dados faltantes do tipo MNAR. O R conta ainda com outros pacotes para realização de imputação (Amelia, mitools, mix, pan, etc).

Após as análises, a *performance* dos procedimentos de simulação e imputação foi avaliada utilizando a média Monte Carlo dos coeficientes e as variâncias determinadas para cada cenário – os indicadores utilizados estão expostos na seção seguinte.

4.5.3 Avaliação do procedimento de IM – indicadores de *performance*

Para avaliar a qualidade do procedimento de imputação múltipla, foram utilizados indicadores de *performance*, para se verificar a acurácia e dispersão das estimativas obtidas com os dados imputados. Alguns indicadores foram adaptados para considerar as simulações Monte Carlo realizadas neste estudo, e, tais modificações são semelhantes às conduzidas por Zhou e colaboradores (2001).

O desvio médio quadrático foi usado para estimar a média geral do erro do método de IM utilizado em cada cenário após a replicação (equação 4.2) (Junger, 2008).

O viés foi calculado como a média das diferenças entre os valores originais e os imputados após combinação dos resultados das replicações (equação 4.3).

Para avaliação da dispersão das simulações, a variância dos valores imputados e combinados em cada replicação foi calculada.

$$DMQ = \frac{1}{m}\sqrt{\sum_{i=1}^{m}\left(Q_i - \widehat{Q}_i\right)^2} \quad (4.2)$$

$$VIÉS = \frac{1}{m}\sum_{i=1}^{m}\left(Q_i - \widehat{Q}_i\right) \quad (4.3)$$

Nas equações anteriores, *m é* o número de replicações utilizadas (100), Q_i os valores originais no banco "completo" e $\widehat{Q}_i$ os valores imputados combinados para cada replicação. Novamente, toda a avaliação de *performance* foi feita utilizando o programa R (R Development Core Team, 2012), versão 2.15.

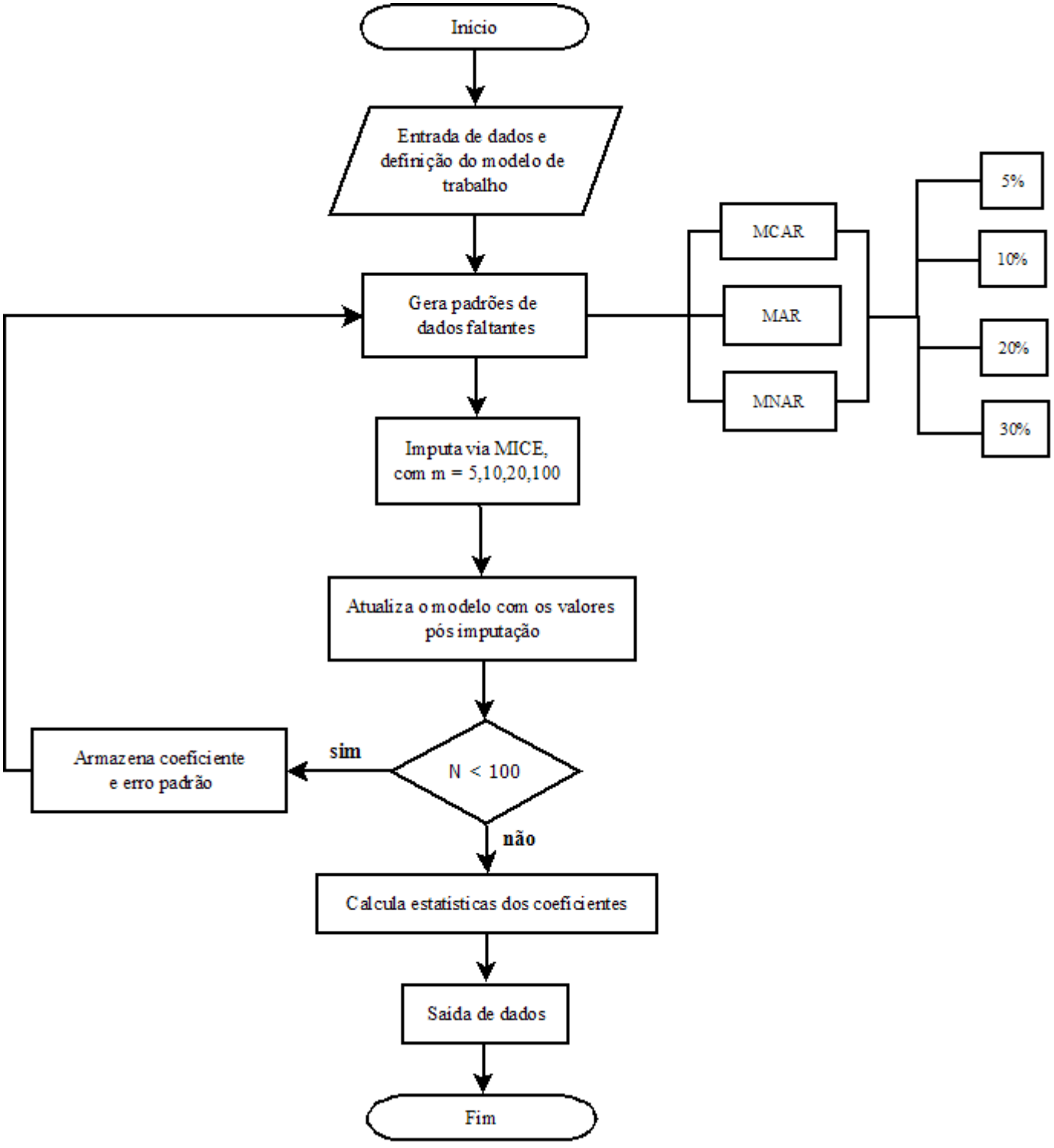

Figura 3 – Algoritmo de simulação e imputação aplicado aos dados de 1505 mulheres participantes da linha de base (Fases 1 e 2) do Estudo Pró-Saúde – RJ (1999-2001).

5 **RESULTADOS**

5.1 **Artigo 1**

Imputação múltipla de dados faltantes em análise de sobrevivência: ***Aplicação no Estudo Pró-Saúde.***

Introdução

Dados faltantes são um problema comum em estudos epidemiológicos, e, dependendo da forma como ocorrem, ignorar a não-resposta pode comprometer a inferência a ser feita a partir dos dados disponíveis (Haukoos & Newgard, 2007; Rothman *et al.*, 2008).

Existem diversas maneiras de se lidar com os dados faltantes. Estas variam desde o descarte de indivíduos com alguma informação faltante – a chamada análise de observações completas, *default* dos aplicativos de análise estatística (Moons *et al.*, 2006) - a métodos mais complexos, que envolvem o "resgate" do que está faltando com base nas informações disponíveis (Little & Rubin, 1987; Donders *et al.*, 2006).

Algumas destas técnicas disponíveis exigem reflexão principalmente quanto ao mecanismo de ocorrência dos dados faltantes e, em determinadas situações, quanto ao padrão desta. Quanto aos mecanismos, estes podem ser: dado faltante completamente aleatório (em inglês, *missing completely at random* – MCAR): ocorre quando a informação faltante é independente da variável de exposição e do desfecho, isto é, os dados observados constituem uma subamostra aleatória dos dados totais; dado faltante aleatório (*missing at random* – MAR): ocorre quando a informação faltante depende apenas do que foi observado e não do que não foi, e, tal informação pode ser reconstituída com base nos dados disponíveis; e dado faltante não aleatório (*missing not at random* – MNAR): ocorre quando a informação faltante depende do que não foi observado e, possivelmente, daquilo que foi de fato observado (Little & Rubin, 1987). Com base no pressuposto de os dados faltantes serem do tipo MAR, as técnicas de imputação foram desenvolvidas (Rubin, 1987; Donders *et al.*, 2006).

Quanto ao padrão de ocorrência dos dados faltantes, este pode ser monotônico ou não-monotônico. No primeiro, comum em estudos longitudinais, blocos de variáveis apresentam cada vez mais dados faltantes, ou de acordo com as ondas de seguimento ou de acordo com a sequência de variáveis em uma mesma onda. No segundo, tal padrão crescente de não-

resposta não é observado (Little & Rubin, 1987). Cabe ressaltar que a não-resposta pode ocorrer em uma unidade inteira ou em alguns itens de uma mesma unidade.

Apesar dos diversos métodos disponíveis para se tratar a não-resposta (Little & Rubin, 1987 trazem uma revisão sobre alguns deles), a literatura aponta a imputação múltipla (IM) como a mais adequada em diversas situações (Zhou *et al.*, 2001; Klebanoff & Cole, 2008).

A técnica proposta por Rubin ao final da década de 1970 (Rubin, 1976; Rubin, 1978) é essencialmente bayesiana e envolve: especificação de um modelo paramétrico para os dados completos sob MAR; suposição de uma distribuição *a priori* não informativa para os parâmetros desconhecidos do modelo; e, simulação de múltiplas amostras independentes da distribuição *a posteriori* condicional dos valores faltantes, dada a informação observada (Little & Rubin, 1987; Enders, 2010). A partir da amostragem da distribuição *a posteriori*, são criadas múltiplas cópias do banco de dados (‘*m*’ bancos, sendo $m>1$), com os dados faltantes substituídos pelos valores imputados (amostrados). Sequencialmente, podem ser utilizadas técnicas estatísticas tradicionais e o modelo de interesse deve ser aplicado a cada banco gerado. Para combinação dos resultados dos *m* bancos de dados, as “regras de Rubin” (Rubin, 1987) são aplicadas e os erros-padrão das estimativas passam a incorporar a variância existente em cada banco criado, bem como a variância que existe entre os *m* bancos. Desta forma, a incerteza associada à imputação é levada em conta (Canizares *et al.*, 2004; Sterne *et al.*, 2009).

A forma de se conduzir a imputação múltipla pode ser dividida em dois grandes grupos: “*Joint Modelling* (JM)” e “*Sequential Regression Multiple Imputation*” – ou, como mencionado por outros autores – “*Fully Conditional Specification* (FCS)” (van Buuren *et al.*, 2006; He, 2010).

Neste trabalho, o interesse recai sobre o grupo denominado *Fully Conditional Specification* – FCS. Ao se utilizar desta abordagem, o procedimento geral de IM passa a incorporar modelos condicionais separados para cada variável incompleta. FCS pode ser mais adequada do que os métodos de JM quando não há uma distribuição multivariada apropriada para lidar com o problema que se tem – utilizando-se FCS pode-se facilmente acomodar características complexas dos dados em modelos de regressão construídos de acordo com os critérios aplicados nas análises de dados comuns (van Buuren *et al.*, 2006).

Dentre os métodos de FCS mencionados por van Buuren e Groothuis-Oudshoorn (2011), o procedimento de IM por equações em cadeia (MICE) vem recebendo destaque nos últimos anos (Nunes *et al.*, 2009; White & Royston, 2009; Nunes *et al.*, 2010; Azur *et al.*, 2011). MICE constitui uma alternativa flexível para considerar diferentes distribuições de

diferentes tipos de variáveis, que não podem ser agrupadas sob uma distribuição conjunta única, como nos métodos de JM.

O procedimento geral de imputação por equações em cadeia pode ser resumido em quatro passos (Azur *et al.*, 2011; White *et al.*, 2011):

> 1. Uma imputação única (como imputação por regressão ou pela média) é realizada em cada valor faltante no banco de dados. Esse valor imputado pode ser considerado como um "marcador de posição". Outra maneira de se "completar" inicialmente esses dados é através de uma amostragem aleatória simples dos valores observados, com reposição.
> 2. Os "marcadores de posição" de uma das variáveis – a primeira na sequência dos modelos determinados - (VAR 1) são removidos, e seus valores voltam a ser dados faltantes.
> 3. Os valores observados de "VAR 1" do passo 2 são regredidos nas outras variáveis do modelo de IM, que pode ou não consistir de todas as variáveis no conjunto de dados, dependo das especificações do analista. "VAR 1" seria a variável dependente no modelo de regressão e todas as demais são as co-variáveis deste modelo. Os modelos de imputação operam sob os mesmos pressupostos dos modelos tradicionais.
> 4. Os valores faltantes de "VAR 1" são então substituídos por amostras da distribuição preditiva *a posteriori* desta variável. Esta amostragem ocorre utilizando o amostrador de Gibbs. Na sequência, quando "VAR 1" é utilizada como variável independente nos modelos de imputação de outras variáveis (VAR 2, VAR 3, ... , VAR *K*), tanto os dados observados quanto essas imputações são utilizadas.

Os passos 2-4 são repetidos para cada variável que tenha dado faltante. O ciclo de cada variável constitui uma iteração. No fim de um ciclo, todos os valores faltantes são substituídos com as predições oriundas das regressões que refletem as relações existentes na porção observada dos dados. As imputações são atualizadas e retidas ao fim de cada ciclo, o que gera um banco de dados imputado. A quantidade de ciclos pode ser determinada pelo pesquisador e, em geral, a convergência será alcançada quando os parâmetros que governam as imputações estejam estáveis (White *et al.*, 2011).

A característica mais importante do MICE é sua habilidade em lidar com diferentes tipos de variáveis, cada uma com seu próprio modelo de imputação, que é atualizado com

base no modelo anterior, e utilizado para determinar o modelo posterior. Por esta razão, o procedimento vem sendo utilizado inclusive em estudos que se utilizam de análises de sobrevivência, no quais aparentemente é capaz de produzir resultados satisfatórios (van Buuren *et al.*, 1999; White & Royston, 2009).

Em relação a este tipo de análise, é importante ressaltar que, no que tange à seleção das variáveis a compor o modelo e ao procedimento geral de imputação, a IM ocorre de maneira semelhante às demais análises estatísticas. A particularidade quando da construção dos modelos de imputação em sobrevivência diz respeito à inclusão do desfecho – fundamental para se evitar diluição da associação entre as co-variáveis e o desfecho do modelo de análise (Moons *et al.*, 2006): nos modelos de regressão, deve-se considerar o status (censura ou caso) e o tempo de contribuição de cada participante ao estudo como desfechos, e não só o status do indivíduo. Ainda não há um consenso sobre a forma como o tempo deva ser incorporado ao modelo de imputação: se na forma tradicional, logarítmica ou ambas (White & Royston, 2009).

Desta forma, além das considerações quanto aos pressupostos necessários à adoção de IM, à forma funcional do modelo e as variáveis que o irão compor (van Buuren *et al.*, 1999), à quantidade de bancos de dados a ser gerado (*m*), e, quanto à maneira que a imputação ocorrerá – se via JM ou FCS, a forma como o tempo de sobrevida será incorporado ao modelo constitui a principal distinção da IM em modelos de sobrevivência, exigindo atenção especial do analista.

No presente estudo, espera-se contribuir para o preenchimento desta lacuna identificada na literatura, verificando se a imputação múltipla via MICE é, de fato, eficiente em análises de sobrevivência, e, de que forma o tempo pode ser incorporado nesta análise. Na seção de métodos, são descritas as características gerais do Estudo Pró-Saúde, uma coorte entre funcionários técnico-administrativos de uma universidade do Rio de Janeiro, fonte dos dados utilizados nas análises. São apresentados os cenários de simulação e imputação bem como a avaliação de *performance* destes procedimentos. Nos resultados, os cenários MCAR, MAR e MNAR são demonstrados separadamente, e comparados aos valores obtidos com os dados originais e a análise de observações completas. Os resultados são seguidos pela discussão e conclusão.

Métodos

Para se verificar a eficiência da IM por equações em cadeia em análises de sobrevivência, optou-se por utilizar uma análise deste tipo recentemente concebida do Estudo Pró-Saúde, na qual se tenha identificado um percentual de dados faltantes inferior a 5% dentro dos modelos de análise utilizados. Isto é, a combinação do percentual de não-resposta de cada variável deveria resultar em um percentual de dados faltantes total menor que 5% entre as variáveis explicativas e, com isso, após a remoção dos indivíduos com estas informações não observadas, o banco seria considerado "completo". A adoção de 5% como critério de decisão foi arbitrária e necessária para que se decidisse por um modelo dentre os diversos disponíveis.

A partir do banco de dados utilizado nesta análise de referência, foram conduzidas simulações de não-resposta, que, ao serem imputadas, permitiram comparar a eficiência da imputação múltipla em relação à análise de observações completas, tendo como "padrão-ouro" o banco de dados original, considerado completo.

Cabe ressaltar que o propósito deste trabalho não é debater sobre o modelo teórico-conceitual utilizado ou acerca de associações encontradas ou não. Trata-se de um exemplo de aplicação com fins ilustrativos da técnica de imputação múltipla neste cenário.

Estudo Pró-Saúde – descrição geral

O Estudo Pró-Saúde (EPS) é um estudo longitudinal realizado entre funcionários técnico-administrativos de uma universidade localizada no Rio de Janeiro, cujo início se deu em 1998. Entre seus objetivos principais figuram a investigação do papel de determinantes sociais no estado de saúde dos indivíduos (Faerstein *et al.*, 2005).

Encontra-se atualmente em sua quarta etapa, sendo as anteriores realizadas nos anos de 1999 (Fase 1), 2001 (Fase 2) e 2006 (Fase 3). Para a realização deste trabalho, foram usadas as informações provenientes da linha de base (Fases 1 e 2) somente. Detalhes sobre a população de estudo encontram-se em Faerstein e colaboradores (2005).

Mesmo com todas as estratégias utilizadas para se evitar perdas ao longo do estudo (Faerstein *et al.*, 2005), ainda se pode observar a ocorrência de não-resposta entre os dados do EPS. Isto decorre principalmente do fato de os questionários serem autopreenchíveis, e, muitas vezes, o participante esquecer de responder alguma pergunta ou a "pular" acidentalmente. Há ainda situações nas quais o participante ignora ou prefere não fornecer as

informações requisitadas no questionário. Outra razão que leva à perda dos dados é a marcação de mais de uma opção de resposta, impossibilitando a identificação de qual delas é a adequada, e, consequentemente, levando ao descarte da informação daquele indivíduo naquela variável. Além disto, a perda entre as fases da coorte também ocorre, já que alguns funcionários se recusam a participar novamente do estudo por diversas justificativas.

A coleta de dados do EPS ocorre com o auxílio de pesquisadores treinados e supervisores, no local de trabalho do indivíduo, por meio de questionários autopreenchíveis e aferições de peso, estatura, circunferência abdominal e pressão arterial, utilizando métodos padronizados e com controle de qualidade regular (Faerstein *et al.*, 2005).

Nas duas etapas da linha de base foram realizados pré-testes a fim de se avaliar a estrutura e adequação do questionário bem como aspectos referentes ao processamento de dados. Além disto, ocorreram estudos-piloto entre os funcionários que não pertenciam ao quadro efetivo da unidade, a fim de não apenas testar as etapas de coleta dos dados na linha de base, mas também a confiabilidade dos instrumentos. A repetição da aplicação dos questionários ocorria em um intervalo de duas semanas (Faerstein *et al.*, 2005; Boclin, 2011). O EPS foi aprovado pelo Comitê de Ética em Pesquisa da instituição na qual é conduzido.

Estudo de referência

Selecionou-se a análise realizada por Boclin (2011), em tese de doutorado recentemente defendida. Uma das propostas de tal trabalho foi averiguar se posição sócio-econômica durante a infância, início da vida adulta, ou ao longo da vida era mediadora da relação entre cor/raça e a ocorrência de miomas uterinos (MU) na população. Para conduzir tal investigação, utilizou-se os dados da linha de base do EPS para a população feminina, que contava com 1819 participantes, dentre 2466 funcionárias elegíveis (73,8% das elegíveis foram avaliadas em 1999-2001).

O desfecho utilizado na análise foi o diagnóstico médico auto-relatado de miomas uterinos – avaliado por meio da questão: "*Alguma vez um médico lhe informou que você tinha mioma uterino, um tumor benigno no útero?*", cujas respostas possíveis eram "Sim" e "Não", tratando-se, portanto, de variável dicotômica.

Outras informações coletadas acerca dos miomas uterinos foram a idade da participante quando do diagnóstico do mioma e sobre a realização de cirurgia de retirada do

útero (histerectomia) e a idade da mulher quando submetida a esta, para que se reconstituísse adequadamente o período de seguimento das participantes.

A cor/raça das participantes, utilizada como variável de exposição principal, foi coletada através da pergunta aberta "*Em sua opinião, qual é a sua cor ou raça?*". Após agrupamento dos relatos das participantes, a variável em questão foi categorizada neste estudo da seguinte maneira: branca, parda, preta e amarela. A categoria amarela foi excluída das análises por apresentar um número pequeno de participantes (n=8, 0,44%), o que poderia introduzir 'ruídos' às análises.

Outras variáveis utilizadas no modelo como marcadoras de acesso e utilização de serviços de saúde foram: Plano de saúde; Realização de teste Papanicolaou; Realização de exame de mama.

A variável "plano de saúde" foi coletada em três categorias: 1 - Sim, como titular; 2 - Sim, como dependente; 3 – Não, e, posteriormente, dicotomizada. As variáveis "realização de teste Papanicolaou" e "realização de exame de mama" foram coletadas em quatro categorias: 1 – Nunca fiz o exame; 2 – Há mais de 3 anos; 3 – Entre 1 e 3 anos atrás; 4 – há menos de 1 ano. As duas também foram dicotomizadas em 'nunca realizou ou realizou há mais de 3 anos' e "realizou há menos de 3 anos" para o presente trabalho.

A escolaridade da participante também fazia parte dos modelos em questão, tendo sido coletada em sete categorias e, posteriormente, transformada em três: 0 - até 1º grau completo; 1 – até 2º grau completo e 2 – universitário completo ou mais.

Outra variável importante foi a idade da participante, que compunha os modelos de análise de forma contínua e, neste trabalho, para a simulação e imputação, foi inserida na forma de tercis, sobretudo por questões práticas quando da simulação.

Ainda na simulação, a variável referente à cor/raça de acordo com a classificação proposta pelo IBGE, coletada através da pergunta *"O Censo Brasileiro (IBGE) usa os termos preta, parda, branca, amarela e indígena para classificar a cor ou raça das pessoas. Se você tivesse que responder ao Censo do IBGE hoje, como se classificaria a respeito de sua cor ou raça?"*, cujas respostas possíveis eram: preta, parda, branca, amarela e indígena, também foi utilizada.

Para lidar com o problema de os dados terem sido coletados transversalmente (dados da linha de base do EPS), buscou-se resgatar as histórias de seguimento das participantes, a partir das informações por elas relatadas. Assumiu-se que o seguimento se iniciou aos 20 anos de idade para todas as mulheres e o tempo final de "acompanhamento" delas foi determinado da seguinte maneira: para mulheres que não apresentavam miomas uterinos, utilizou-se a

idade em 1999; para as mulheres que relataram miomas, a idade do diagnóstico do mesmo, e, para mulheres que relataram histerectomia, a idade da cirurgia. Determinou-se como tempo final de seguimento máximo a idade de 50 anos. Desta forma, dez casos de miomas diagnosticados após esse período foram censurados.

Nas análises multivariadas, foram utilizados modelos de riscos proporcionais de Cox para estimar a Razão de Hazards com intervalos de 95% de confiança (IC 95%) (Boclin, 2011). O modelo utilizado neste trabalho é o que tem por desfecho o status da participante (caso - 1; censura – 0) e o tempo de seguimento conforme os critérios antes mencionados. Como variáveis de exposição estão a cor/raça e a idade das participantes e como fatores de confusão do modelo, as variáveis marcadoras de acesso aos serviços de saúde (plano de saúde, exame de Papanicolaou, exame de mama) e a escolaridade da mulher.

Organização do banco de dados

Utilizando o mesmo conjunto de dados analisado por Boclin (2011), procedeu-se a exclusão das participantes consideradas inelegíveis para este estudo: aquelas com informação faltante no desfecho sob investigação – por se tratar de um modelo de sobrevivência, foram excluídas mulheres sem informação sobre a ocorrência de mioma uterino ou sem informação sobre a idade do diagnóstico do mesmo. Além disto, optou-se por excluir as mulheres que afirmaram ter realizado histerectomia, mas não informaram a idade da mesma e aquelas que não informaram sobre realização ou não desta cirurgia, já que não se podia determinar a extensão do seu período sob risco. Em seguida, foram removidas as mulheres cujo mioma uterino ou a histerectomia ocorreram antes dos 20 anos de idade, já que haviam sido consideradas inelegíveis no estudo de base. Por fim, mulheres de cor/raça amarela foram excluídas.

Com as exclusões acima mencionadas, o banco de dados inicial, com 1819 mulheres, passou a ser constituído de 1593 participantes. Destas, 5,5 % (n=88) não tinham informação em alguma das variáveis do modelo proposto ou outras variáveis a serem utilizadas durante os procedimentos de simulação ou imputação. Com a remoção destas últimas, o banco de dados considerado completo (submetido à simulação nesta dissertação) consistiu de 1505 mulheres, todas com informações completas no desfecho, idade, cor/raça, plano de saúde, realização de exame de mama e Papanicolaou, escolaridade e cor/raça de acordo com a classificação proposta pelo IBGE.

Este banco de dados foi organizado de forma a permitir uma análise utilizando o modelo de Cox. Toda a montagem do banco de dados foi feita utilizando o *software* R (R Development Core Team, 2012), versão 2.15.

Procedimento de simulação

Para este estudo, o percentual de dados faltantes do modelo utilizado (aproximadamente 5% após a combinação das variáveis explicativas) permitiu a obtenção de um banco de dados com informação completa nas co-variáveis de interesse e no desfecho. Este banco foi assumido como referência e analisado através do modelo de riscos proporcionais de Cox. Os resultados desta análise são considerados "padrão-ouro" quando da comparação com os resultados obtidos na IM.

Na sequência foi simulada não-resposta nestes dados obedecendo ao algoritmo exposto na figura 1. Fixou-se o padrão de dados faltantes em não-monotônico, já que esse era o padrão observado entre os dados antes da remoção dos 5,5% faltantes. A não-resposta foi simulada na variável cor/raça das participantes (variável multinomial).

Ocorreu variação no percentual de dados faltantes – de um cenário com menor percentual (5%) a valores mais elevados (10, 20, e 30%); no mecanismo de ocorrência dos dados faltantes – MCAR, MAR e MNAR; e, quando da imputação, houve variação na quantidade de banco de dados (*m*) a ser criado com o procedimento - m= 5, 10, 20 e 100. Considerou-se suficiente m=100, já que para alcançar uma eficiência relativa de 100%, seria necessário $m \to \infty$, de acordo com a equação proposta por Rubin (1987) e um *m* elevado levaria a uma situação computacionalmente intensiva.

Para determinar quais indivíduos deveriam ser transformados em dados faltantes em cada cenário, determinou-se um modelo logístico para predição da probabilidade da ocorrência de dados faltantes na variável cor/raça (ou seja, desfecho: 1 – ser dado faltante para cor/raça; 0 – não o ser), utilizando o banco de dados com n=1593 participantes, que possuía 3,2% de *missing* na mesma. Um procedimento semelhante é descrito por Zhou e colaboradores (2001).

O modelo logístico no cenário MAR contou com as variáveis idade em tercil e escolaridade das participantes (em três categorias), já que pareciam capazes de explicar, em parte, a ocorrência do desfecho em questão (ou seja, ser ou não dado faltante na variável cor/raça). Com isto, havia 9 padrões possíveis de co-variáveis, e, dentro de cada um eles,

determinou-se a probabilidade de ocorrência de dados faltantes para cor/raça. Estas probabilidades associadas a cada padrão foram então aplicadas aos mesmos padrões no banco de dados completo (n=1505) e, multiplicando-se estes valores pela quantidade de participantes em cada padrão, pôde-se determinar quantos indivíduos deveriam ser removidos em cada um deles, a fim de se gerar 5, 10, 20 e 30% de dados faltantes na totalidade.

Para o cenário MNAR, além de idade e escolaridade, incluiu-se a variável cor/raça de acordo com a classificação do IBGE, a fim de se forçar a associação entre o desfecho (ser ou não dado faltante) e a (não) resposta à questão sobre sua cor/raça. Isto levou à existência de 18 padrões de co-variáveis, que tiveram suas probabilidades de ocorrência do desfecho calculadas de forma semelhante ao cenário anterior, e, novamente, foram multiplicadas pela quantidade de indivíduos em cada padrão e recalculadas para gerar 5, 10, 20 e 30% de não-resposta.

Para MCAR, foi realizada uma amostragem aleatória utilizando os números de identificação das participantes, obedecendo as quantidades necessárias para gerar 5% (n=76), 10% (n=152), 20% (n=301) e 30% (n=452) de não-resposta. Os indivíduos sorteados foram então transformados em dados faltantes na variável cor/raça.

A partir de então, já se conhecendo quantos e quais indivíduos deveriam ter a informação sobre sua cor/raça ‘apagada’ do banco de dados completo, procedeu-se a simulação seguida da imputação de dados.

Procedimento de imputação

Os bancos de dados gerados com as simulações foram imputados pela técnica de IM por equações em cadeia (MICE). O modelo de imputação foi construído conforme a seleção de variáveis proposta por van Buuren e colaboradores (1999), considerando as particularidades de cada variável disponível. Foram incluídas na matriz de predição da cor/raça das participantes as variáveis: idade (em tercis), escolaridade (três categorias), plano de saúde (dicotomizada), realização de exame de mama, realização de exame de Papanicolaou (também dicotomizadas), status da participante (se caso ou censura) e tempo de contribuição no estudo (em anos).

O modelo de análise já havia sido estabelecido *a priori* no estudo de base, e foi repetido em cada banco de dados imputado, utilizando o modelo de riscos proporcionais de Cox da mesma maneira. Para evitar problemas de convergência deste modelo, o número

máximo de iterações foi alterado para 100. Ao final das análises, os valores de cada banco de dados foram combinados, obedecendo as "regras de Rubin" (Rubin, 1987).

Para determinar a distribuição das estimativas obtidas em cada cenário de imputação, o processo foi replicado 100 vezes e, ao final, foi determinada a média Monte Carlo dos coeficientes do modelo de regressão para a variável de interesse (cor/raça) e seus erros-padrão, bem como a variância entre os coeficientes.

Utilizando as médias dos coeficientes e erros, determinou-se a Razão de Hazards e seu intervalo de confiança (IC95%) em cada cenário, apenas para a variável de exposição principal, a fim de compará-los com os valores do "padrão-ouro" e com os da análise de observações completas. Estas últimas também foram replicadas 100 vezes, a fim de se determinar a distribuição das estimativas desta análise, após simulação dos dados faltantes, da mesma forma como ocorreu com a IM.

A análise foi conduzida no programa estatístico R (R Development Core Team R, 2012), versão 2.15, utilizando as bibliotecas *MICE* (van Buuren & Groothuis-Oudshoorn, 2011) e *SURVIVAL* (Therneau, 2012).

Avaliação de *performance*

Para avaliar a qualidade do procedimento de simulação e imputação múltipla, foram utilizados indicadores de *performance*, para se verificar a acurácia e dispersão das estimativas obtidas com os dados imputados. Alguns indicadores foram adaptados para considerar as simulações Monte Carlo realizadas neste estudo, considerando o trabalho de Zhou e colaboradores (2001).

O desvio médio quadrático foi usado para estimar a média geral do erro do método de IM utilizado em cada cenário após a replicação (equação 1) (Junger, 2008).

$$DMQ = \frac{1}{m}\sqrt{\sum_{i=1}^{m}\left(Q_i - \widehat{Q}_i\right)^2} \quad (1)$$

O viés foi calculado como a média das diferenças entre os valores originais e os imputados após combinação dos resultados das replicações (equação 2).

$$VIÉS = \frac{1}{m}\sum_{i=1}^{m}\left(Q_i - \hat{Q}_i\right) \quad (2)$$

Para avaliação da dispersão das simulações, a variância dos valores imputados e combinados em cada replicação foi calculada.

Nas equações anteriores, *m é* o número de replicações utilizadas (100), Q_i os valores originais no banco "completo" e $\hat{Q}_i$ os valores imputados combinados para cada replicação. Novamente, toda a avaliação de *performance* foi feita utilizando o programa R (R Development Core Team, 2012), versão 2.15.

Resultados

A análise do banco de dados completo revelou os seguintes valores de RH (e IC95%): para mulheres de cor/raça branca (linha de base) – RH=1,0; para pardas, RH= 1,1365 (0,8534 – 1,5137); e, para pretas, RH= 1,7925 (1,3805 – 2,3276). Estes valores são os considerados 'padrão-ouro' para se verificar a eficiência da IM em produzir estimativas não viesadas. Tais valores são compatíveis com os valores de RH encontrados no mesmo modelo de análise do estudo de base (Boclin, 2011), com uma ligeira discrepância para pretas - no estudo de referência, RH= 1,7 (IC95% 1,3 - 2,3) e maior diferença para pardas – RH do estudo de base = 1,2 (0,9 – 1,6). Vale ressaltar que, neste estudo de referência, o procedimento adotado foi o de análise de observações completas.

Ao se adotar este mesmo procedimento nos bancos de dados simulados deste trabalho, encontrou-se os valores expostos na tabela 1. Observa-se que, independente do mecanismo de simulação de não-resposta, quanto maior o percentual de dados faltantes, mais distantes dos valores 'verdadeiros' estão as estimativas. Além disto, e conforme esperado, a amplitude dos intervalos de confiança também aumentam, à medida que o percentual de dados disponíveis diminui.

Os cenários mais discrepantes são MAR e MNAR, ambos nos percentuais 20 e 30. O cenário MCAR só apresentou estimativas mais discrepantes do valor verdadeiro quando o percentual de DF era de 30%. As diferenças parecem ser semelhantes para as categorias parda e preta, exceto no cenário MAR 30%, no qual a disparidade é maior para pardas.

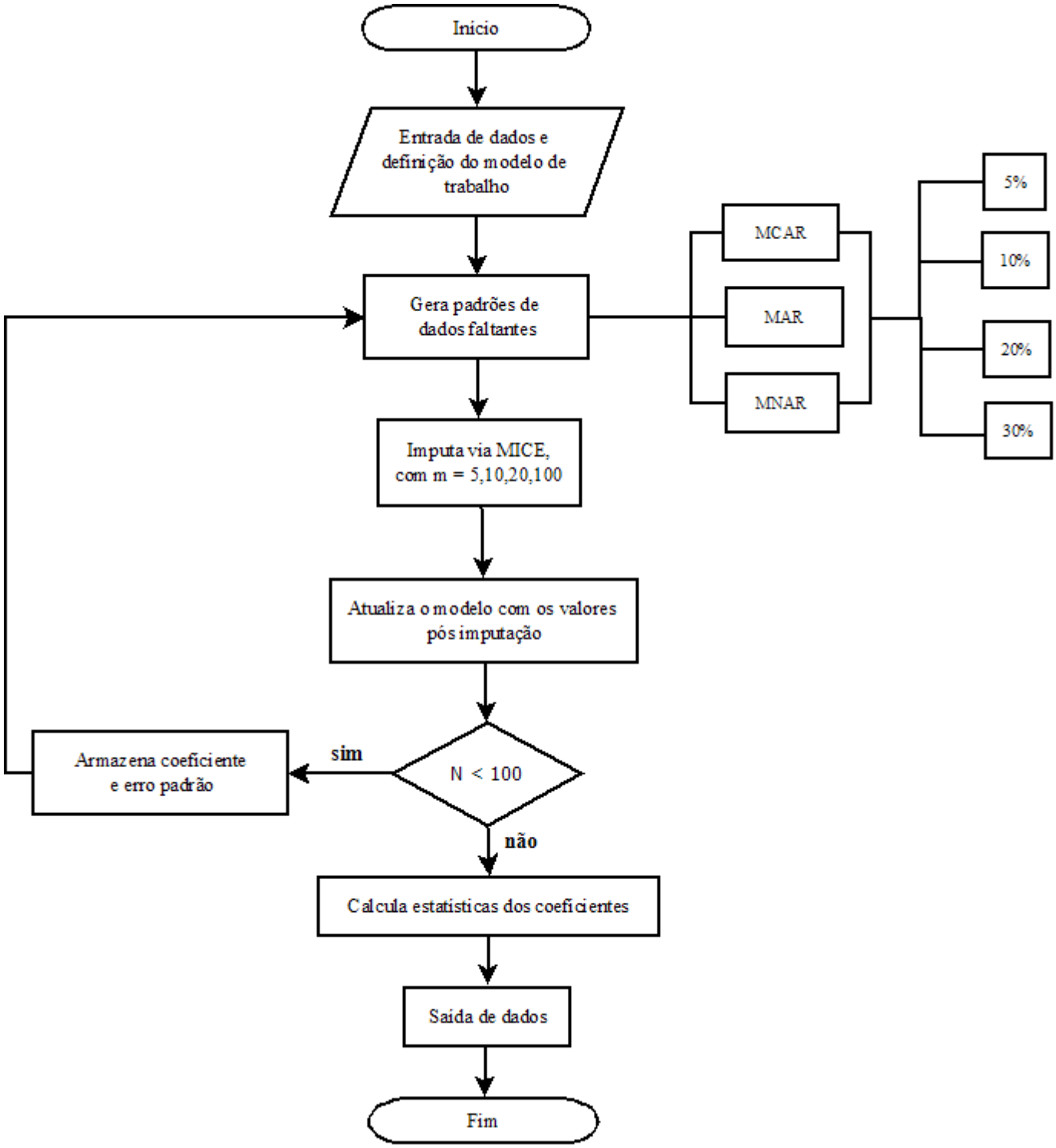

Figura 1 – Algoritmo de simulação e imputação aplicado aos dados de 1505 mulheres participantes da linha de base (Fases 1 e 2) do Estudo Pró-Saúde – RJ (1999-2001).

Para 5% de dados faltantes, independente do mecanismo de não-resposta simulado, as estimativas pontuais obtidas são as mais próximas dos valores originais.

MCAR

A simulação dos dados faltantes no mecanismo MCAR e posterior imputação múltipla dos mesmos teve como resultados os valores expressos na tabela 2. Maiores discrepâncias são encontradas no maior percentual de DF (30%), independente da quantidade de bancos de dados gerada na imputação, porém, com *m* distintos para pretas (pior cenário com m= 10 e m=100) e pardas (pior cenário com m=5 e m=20). Diferenças também ocorrem nos cenários 10% e m=10 para pardas e 20% e m=100 para pretas.

Os cenários nos quais as estimativas pontuais mais se aproximaram do valor verdadeiro foram 20% com m=20 e 10% com m=100 para as duas categorias. Porém, na configuração 20% com m=20, a diferença entre os intervalos de confiança é maior, sendo a amplitude dos intervalos para os valores imputados superior à amplitude dos valores do padrão-ouro tanto para pardas quanto para pretas.

As medidas de *performance* (tabela 5) condizem com as estimativas encontradas. Para o DMQ, observa-se que, para todas as configurações, os valores são baixos e consistentes entre si, indicando que a replicação das imputações fornece bons resultados, independente do *m* adotado.

Em relação ao viés, novamente, os valores são baixos, e, concordando com os resultados da tabela 2, o menor viés ocorre para a configuração 10% com m=100 para a categoria parda. Os maiores valores ocorrem para as configurações 20% com m=100 e 30% com m=100 para pretas e 30% com m=20 para pardas.

As variâncias entre os 100 betas gerados a cada cenário também demonstram valores pequenos, indicando pouca variabilidade entre os betas combinados após a imputação em cada replicação realizada, confirmando a qualidade do procedimento.

Tabela 1 – Valores das Razões de Hazards (e IC95%) da variável cor/raça após análise de observações completas – dados do Estudo Pró-Saúde-RJ (1999-2001).

		Percentuais de dados faltantes*			
		5% (n=1429)	10% (n=1354)	20% (n=1204)	30% (n=1053)
MCAR	Pardas	1,1382 (0,8482 – 1,5275)	1,1301 (0,8348 – 1,5299)	1,1366 (0,8245 – 1,5668)	1,1121 (0,7884 – 1,5689)
	Pretas	1,7961 (1,3738 – 2,3481)	1,7879 (1,3572 – 2,3555)	1,7836 (1,3308 – 2,3904)	1,7757 (1,2990 – 2,4274)
MAR	Pardas	1,1322 (0,8431 – 1,5206)	1,1269 (0,8311 – 1,5281)	1,1102 (0,8004 – 1,5399)	1,0899 (0,7652 – 1,5526)
	Pretas	1,7962 (1,3729 – 2,3499)	1,7841 (1,3520 – 2,3543)	1,8279 (1,3590 – 2,4585)	1,8065 (1,3114 – 2,4885)
MNAR	Pardas	1,1365 (0,8462 – 1,5264)	1,1304 (0,8341 – 1,5319)	1,1193 (0,8076 – 1,5513)	1,1124 (0,7789 – 1,5887)
	Pretas	1,7967 (1,3728 – 2,3516)	1,7928 (1,3581 – 2,3667)	1,8224 (1,3526 – 2,4554)	1,8191 (1,3146 – 2,5173)

* Pode ocorrer pequena variação nos percentuais por conta de aproximações na quantidade de indivíduos removida em cada categoria.

Tabela 2 – Valores da Razão de Hazards (e IC95%) para a variável cor/raça após simulação e imputação no cenário MCAR – dados do Estudo Pró-Saúde – RJ (1999-2001).

		Percentuais de dados faltantes*			
		5%	10%	20%	30%
m=5	Pardas	1,1443 (0,8527 – 1,5357)	1,1436 (0,8452 – 1,5474)	1,1418 (0,8283 – 1,5740)	1,1264 (0,7983 – 1,5894)
	Pretas	1,7976 (1,3743 – 2,3512)	1,7992 (1,3653 – 2,3708)	1,8038 (1,3465 – 2,4165)	1,8104 (1,3239 – 2,4757)
m=10	Pardas	1,1374 (0,8473 – 1,5266)	1,1491 (0,8494 – 1,5544)	1,1444 (0,8305 – 1,5769)	1,1353 (0,8049 – 1,6013)
	Pretas	1,8004 (1,3769 – 2,3541)	1,8099 (1,3738 – 2,3843)	1,8060 (1,3479 – 2,4198)	1,8191 (1,3304 – 2,4881)
m=20	Pardas	1,1461 (0,8544 – 1,5374)	1,1420 (0,8439 – 1,5454)	1,1352 (0,8230 – 1,5658)	1,1172 (0,7919 – 1,5760)
	Pretas	1,7918 (1,3700 – 2,3433)	1,8002 (1,3666 – 2,3714)	1,7977 (1,3423 – 2,4076)	1,7899 (1,3112 – 2,4434)
m=100	Pardas	1,1408 (0,8502 – 1,5308)	1,1356 (0,8390 – 1,5370)	1,1286 (0,8190 – 1,5551)	1,1381 (0,8060 – 1,6069)
	Pretas	1,8037 (1,3798 – 2,3578)	1,7982 (1,3656 – 2,3679)	1,7710 (1,3223 – 2,3720)	1,8208 (1,3302 – 2,4924)

* Pode ocorrer pequena variação nos percentuais por conta de aproximações na quantidade de indivíduos removida em cada categoria.

Tabela 3 – Valores da Razão de Hazards (e IC95%) para a variável cor/raça após simulação e imputação no cenário MAR – dados do Estudo Pró-Saúde – RJ (1999-2001).

		Percentuais de dados faltantes*			
		5%	10%	20%	30%
m=5	Pardas	1,1346 (0,8446 – 1,5242)	1,1325 (0,8351 – 1,5359)	1,1179 (0,8071 – 1,5484)	1,1029 (0,7739 – 1,5718)
	Pretas	1,7991 (1,3749 – 2,3542)	1,8028 (1,3662 – 2,3788)	1,7889 (1,3303 – 2,4056)	1,8049 (1,3091 – 2,4884)
m=10	Pardas	1,1396 (0,8486 – 1,5305)	1,1348 (0,8373 – 1,5379)	1,1040 (0,7966 – 1,5301)	1,1040 (0,7736 – 1,5755)
	Pretas	1,8054 (1,3799 – 2,3620)	1,7905 (1,3568 – 2,3627)	1,7984 (1,3373 – 2,4186)	1,8321 (1,3286 - 2,5263)
m=20	Pardas	1,1309 (0,8416 – 1,5196)	1,1384 (0,8397 – 1,5433)	1,1229 (0,8111 – 1,5547)	1,1067 (0,7759 – 1,5784)
	Pretas	1,7982 (1,3742 – 2,3530)	1,8117 (1,3736 – 2,3894)	1,8054 (1,3428 – 2,4273)	1,8270 (1,3244 – 2,5204)
m=100	Pardas	1,1393 (0,8485 – 1,5299)	1,1266 (0,8311 – 1,5272)	1,1126 (0,8022 – 1,5430)	1,0901 (0,7645 – 1,5545)
	Pretas	1,7986 (1,3748 – 2,3531)	1,8000 (1,3645 – 2,3746)	1,8041 (1,3408 – 2,4274)	1,8146 (1,3174 – 2,4993)

* Pode ocorrer pequena variação nos percentuais por conta de aproximações na quantidade de indivíduos removida em cada categoria.

Tabela 4 – Valores da Razão de Hazards (e IC95%) para a variável cor/raça após simulação e imputação no cenário MNAR – dados do Estudo Pró-Saúde – RJ (1999-2001).

		Percentuais de dados faltantes*			
		5%	10%	20%	30%
m=5	Pardas	1,1334 (0,8438 – 1,5224)	1,1394 (0,8412 – 1,5435)	1,1313 (0,8168 – 1,5668)	1,1159 (0,7836 – 1,5892)
	Pretas	1,7948 (1,3710 – 2,3497)	1,7973 (1,3618 – 2,3719)	1,7984 (1,3331 – 2,4260)	1,8326 (1,3249 – 2,5347)
m=10	Pardas	1,1327 (0,8426 – 1,5226)	1,1266 (0,8306 – 1,5281)	1,1192 (0,8074 – 1,5515)	1,1153 (0,7831 – 1,5883)
	Pretas	1,8019 (1,3764 – 2,3590)	1,8016 (1,3647 – 2,3784)	1,8039 (1,3382 – 2,4318)	1,8363 (1,3277 – 2,5397)
m=20	Pardas	1,1311 (0,8422 – 1,5190)	1,1337 (0,8362 – 1,5370)	1,1312 (0,8163 – 1,5675)	1,1064 (0,7764 – 1,5767)
	Pretas	1,7881 (1,3661 – 2,3404)	1,7986 (1,3621 – 2,3749)	1,8214 (1,3507 – 2,4560)	1,8515 (1,3397 – 2,5588)
m=100	Pardas	1,1379 (0,8475 – 1,5278)	1,1427 (0,8435 – 1,5480)	1,1052 (0,7962 – 1,5343)	1,1040 (0,7741 – 1,5746)
	Pretas	1,7939 (1,3700 – 2,3489)	1,8093 (1,3708 – 2,3881)	1,8040 (1,3375 – 2,4330)	1,8098 (1,3087 – 2,5028)

* Pode ocorrer pequena variação nos percentuais por conta de aproximações na quantidade de indivíduos removida em cada categoria.

MAR

No mecanismo MAR, uma situação semelhante à que ocorre em MCAR é encontrada (tabela 3). Novamente, o cenário mais divergente é o percentual de DF 30%, independente do *m* utilizado. Porém, diferentemente do que se observa anteriormente, para pardas, as diferenças nas estimativas começam a aparecer de maneira mais marcante quando o percentual de DF é 20. Para pretas, as maiores discrepâncias aparecem nesta configuração apenas para m=20 e m=100.

O cenário no qual as estimativas de pardas e pretas são mais próximas dos valores verdadeiros foram 5% com m=20 e 10% com m=10, com pequenas diferenças nos limites dos intervalos de confiança para ambos.

A avaliação de *performance* deste mecanismo (tabela 6) indicou valores baixos para todas as medidas, conforme ocorreu em MCAR. Para o DMQ, os valores baixos concordam com o esperado para este mecanismo.

Em relação ao viés, os valores mais elevados, ocorrem quando o percentual de dados faltantes é de 30%, para todos os m, na categoria parda. O menor viés ocorre para 10% com m=10 tanto para pardas (Viés=0,0015) quanto para pretas (Viés=0,0011). Para a variância, todos os valores são baixos e condizentes com o esperado após as replicações.

MNAR

Quanto ao cenário MNAR, as estimativas de RH mais próximas dos valores verdadeiros ocorrem no cenário 5% para todas as variações de *m*.

Estimativas mais distantes do valor verdadeiro ocorrem no cenário 30% com m=5, m=10 e m=20 para pretas. Na configuração 30% com m=20, a maior diferença de estimativa entre todos os cenários simulados e imputados ocorre para esta categoria. Também há discrepância quando o percentual é de 20% com m=20. Para pardas, as configurações mais divergentes foram: 20% com m=100, 30% com m=20 e 30% com m=100.

Nota-se que, apesar de não ser o cenário com maior quantidade de estimativas discrepantes dos valores verdadeiros (MAR apresenta cenários inteiros com alguma diferença), o mecanismo MNAR, como esperado, apresenta as maiores diferenças quando se compara a estimativa pontual imputada com a verdadeira, tanto para pardas como para pretas.

Na avaliação de *performance* (tabela 7), este mecanismo também apresentou resultados condizentes com o esperado, como ocorreu nos demais mecanismos.

Para DMQ, os valores foram pequenos, indicando pequeno desvio do valor original. O maior valor ocorreu para 30% com m=100 entre as pardas (0,0111). Para variância, apesar de a dispersão entre os betas replicados ser pequena, para 30% de dados faltantes ocorreram os maiores valores, independente do *m* e da categoria de análise.

No que diz respeito ao viés, como ocorre com a variância, os valores mais discrepantes estão no percentual 30 para todos os m, tanto para pardas quanto para pretas. Valores mais baixos de viés ocorrem nos menores percentuais de dados faltantes (5% e 10%) e menor viés ocorre para 5% com m=100 para pretas.

Discussão

No presente trabalho, os resultados encontrados tanto com a análise de observações completas quanto com a IM se mostraram semelhantes aos valores obtidos na análise do banco de dados completo.

A análise de observações completas não apresentou vieses importantes que comprometessem as estimativas pontuais obtidas. A mudança mais significativa nesta situação foi o aumento da amplitude dos intervalos de confiança com o aumento do percentual de dados faltantes, fato já documentado na literatura (Little, 1992; Greenland & Finkle, 1995). Neste estudo, a análise de observações completas pode não ter gerado intervalos de confiança tão amplos por conta do número de observações relativamente alto (n>1000 em todas os percentuais de dados faltantes), o que fez com que mesmo uma perda mais elevada (30%) não comprometesse o poder estatístico de forma significativa (Bono *et al.*, 2007).

No que diz respeito à IM, os resultados condizem com os valores verdadeiros obtidos no banco completo, mesmo quando o mecanismo de ocorrência dos dados faltantes era sabidamente não aleatório. Apesar de este tipo de mecanismo não ser o ideal para aplicação da imputação múltipla, ela foi capaz de fornecer resultados satisfatórios quando comparados ao padrão-ouro.

Tabela 5 – Indicadores de *performance* do procedimento de simulação e imputação de dados no cenário MCAR – Estudo Pró-Saúde-RJ (1999-2001).

			Percentuais de dados faltantes			
			5%	10%	20%	30%
DMQ	m=5	Pardas	0,0031	0,0046	0,0072	0,0097
		Pretas	0,0031	0,0044	0,0071	0,0085
	m=10	Pardas	0,0032	0,0053	0,0070	0,0101
		Pretas	0,0035	0,0047	0,0064	0,0099
	m=20	Pardas	0,0034	0,0049	0,0069	0,0099
		Pretas	0,0034	0,0049	0,0060	0,0087
	m=100	Pardas	0,0035	0,0047	0,0080	0,0095
		Pretas	0,0031	0,0040	0,0072	0,0086
Viés	m=5	Pardas	-0,0068	-0,0062	-0,0046	0,0089
		Pretas	-0,0028	-0,0037	0,0063	-0,0099
	m=10	Pardas	-0,0007	-0,0109	-0,0069	0,0011
		Pretas	-0,0044	-0,0096	-0,0075	-0,0147
	m=20	Pardas	-0,0084	-0,0047	0,0012	0,0172
		Pretas	0,0004	-0,0043	-0,0029	0,0014
	m=100	Pardas	-0,0038	0,0008	0,0070	-0,0013

		Pretas	-0,0062	-0,0032	0,0120	-0,0157
Variância	m=5	Pardas	0,0009	0,0021	0,0053	0,0093
		Pretas	0,0010	0,0020	0,0051	0,0072
	m=10	Pardas	0,0010	0,0028	0,0049	0,0104
		Pretas	0,0012	0,0021	0,0041	0,0099
	m=20	Pardas	0,0011	0,0024	0,0049	0,0096
		Pretas	0,0012	0,0024	0,0037	0,0076
	m=100	Pardas	0,0012	0,0023	0,0064	0,0092
		Pretas	0,0009	0,0016	0,0052	0,0073

Tabela 6 – Indicadores de *performance* do procedimento de simulação e imputação de dados no cenário MAR – Estudo Pró-Saúde-RJ (1999-2001).

			Percentuais de dados faltantes			
			5%	10%	20%	30%
DMQ	m=5	Pardas	0,0034	0,0047	0,0084	0,0094
		Pretas	0,0036	0,0044	0,0080	0,0083
	m=10	Pardas	0,0036	0,0056	0,0074	0,0091
		Pretas	0,0031	0,0052	0,0067	0,0094
	m=20	Pardas	0,0039	0,0051	0,0071	0,0111
		Pretas	0,0031	0,0046	0,0075	0,0102
	m=100	Pardas	0,0032	0,0052	0,0085	0,0096
		Pretas	0,0035	0,0044	0,0077	0,0089
Viés	m=5	Pardas	0,0017	0,0035	0,0165	0,0300
		Pretas	-0,0037	-0,0057	0,0020	-0,0069
	m=10	Pardas	-0,0027	0,0015	0,0290	0,0290
		Pretas	-0,0072	0,0011	-0,0033	-0,0218
	m=20	Pardas	0,0050	-0,0016	0,0120	0,0266
		Pretas	-0,0032	-0,0106	-0,0071	-0,0190
	m=100	Pardas	-0,0024	0,0088	0,0213	0,0417

		Pretas	-0,0034	-0,0042	-0,0064	-0,0122
Variância	m=5	Pardas	0,0011	0,0022	0,0068	0,0081
		Pretas	0,0013	0,0019	0,0065	0,0069
	m=10	Pardas	0,0013	0,0032	0,0046	0,0076
		Pretas	0,0009	0,0028	0,0045	0,0084
	m=20	Pardas	0,0015	0,0027	0,0050	0,0118
		Pretas	0,0010	0,0020	0,0056	0,0102
	m=100	Pardas	0,0010	0,0027	0,0068	0,0076
		Pretas	0,0012	0,0019	0,0060	0,0078

Tabela 7 – Indicadores de *performance* do procedimento de simulação e imputação de dados no cenário MNAR – Estudo Pró-Saúde-RJ (1999-2001).

			Percentuais de dados faltantes			
			5%	10%	20%	30%
DMQ	m=5	Pardas	0,0031	0,0043	0,0072	0,0092
		Pretas	0,0036	0,0047	0,0077	0,0093
	m=10	Pardas	0,0037	0,0047	0,0075	0,0099
		Pretas	0,0034	0,0049	0,0068	0,0087
	m=20	Pardas	0,0028	0,0050	0,0074	0,0090
		Pretas	0,0030	0,0044	0,0070	0,0097
	m=100	Pardas	0,0035	0,0044	0,0072	0,0111
		Pretas	0,0037	0,0044	0,0080	0,0089
Viés	m=5	Pardas	0,0028	-0,0025	0,0046	0,0183
		Pretas	-0,0013	-0,0026	-0,0033	-0,0221
	m=10	Pardas	0,0034	0,0088	0,0154	0,0189
		Pretas	-0,0052	-0,0051	-0,0064	-0,0241
	m=20	Pardas	0,0048	0,0025	0,0047	0,0268
		Pretas	0,0025	-0,0034	-0,0160	-0,0324
	m=100	Pardas	-0,0012	-0,0054	0,0279	0,0290

		Pretas	-0,0008	-0,0093	-0,0064	-0,0096
Variância	m=5	Pardas	0,0010	0,0018	0,0053	0,0082
		Pretas	0,0013	0,0023	0,0060	0,0083
	m=10	Pardas	0,0014	0,0022	0,0054	0,0097
		Pretas	0,0011	0,0024	0,0046	0,0070
	m=20	Pardas	0,0008	0,0026	0,0056	0,0075
		Pretas	0,0009	0,0019	0,0047	0,0085
	m=100	Pardas	0,0012	0,0019	0,0045	0,0115
		Pretas	0,0014	0,0018	0,0065	0,0078

Mesmo com as pequenas diferenças observadas, em nenhum dos cenários aqui avaliados os resultados foram tão discrepantes que fizessem com que os RH significantes o deixassem de ser ou o contrário. As estimativas pontuais e os intervalos de confiança obtidos apresentaram pequenas variações, mas nenhuma delas compromete a aplicação da imputação em qualquer um dos cenários apresentados.

Outros autores (Catellier *et al.*, 2005), ao avaliarem resultados imputados considerando os três mecanismos de não-resposta, apontam que a imputação é efetiva quando os dados faltantes são do tipo MAR ou MCAR, e as estimativas podem estar enviesadas quando MNAR ocorrem. Corroborando os resultados do presente trabalho, Schafer e Graham (2002) mencionam que apesar de se adotar a suposição de que o mecanismo de ocorrência dos dados faltantes seja MAR, por este ser capaz de produzir melhores resultados, os resultados a serem obtidos sob MNAR podem ser igualmente não viesados. Collins e colaboradores (2001) encontraram resultados satisfatórios na imputação múltipla mesmo quando MNAR ocorria, a exemplo dos achados aqui documentados.

Em relação à incorporação do tempo, questão fundamental nos estudos que envolvem análise de sobrevivência, neste trabalho, acrescentar o tempo em anos, sem considerar transformações do mesmo, foi capaz de produzir estimativas consistentes com as do padrão-ouro. Acredita-se que esta incorporação deva ser avaliada em cada situação particular, mas, no cenário aqui apresentado, incorporar o status da participante e seu tempo de participação no estudo em anos foi suficiente para reter a associação entre a exposição principal e o desfecho avaliado, produzindo estimativas de RH semelhantes às esperadas e mantendo a relação existente entre a cor/raça e a ocorrência de miomas uterinos já evidenciada no estudo de referência (Boclin, 2011).

Um ponto positivo a ser ressaltado foi a replicação das simulações e imputações para se criar uma distribuição dos coeficientes combinados ao final da análise dos bancos imputados. Este procedimento fez com que os coeficientes, e, consequentemente, os RH, se tornassem bastante homogêneos e permitiu que os resultados encontrados com as imputações fossem consistentes em todos os cenários. A replicação também pode ter favorecido as estimativas obtidas com a análise de observações completas. Outros estudos também fizeram uso de replicações e encontraram bons resultados com a imputação múltipla (Collins *et al.*, 2001; Zhou *et al.*, 2001).

A avaliação de *performance* dos procedimentos, de maneira geral, corroborou os resultados não viesados obtidos nas replicações. Os baixos valores de DMQ e viés indicam a qualidade das técnicas aplicadas, principalmente por conta da repetição empregada. Já a variância também baixa aponta a pequena dispersão existente entre as 100 replicações realizadas em cada cenário, indicando que, aparentemente, em nenhum deles os coeficientes combinados da imputação apresentaram problemas significativos capazes de afetar a média geral dos mesmos. Observando os intervalos de confiança de cada cenário, nota-se que todos incluem o valor verdadeiro da Razão de Hazards encontrada na análise do banco 'completo', reforçando, novamente, a habilidade do procedimento de imputação múltipla em produzir estimativas semelhantes às esperadas.

Outras medidas de *performance* são sugeridas para se avaliar a qualidade da imputação, mas estas, em geral, são aplicadas em situações nas quais a técnica é conduzida apenas uma vez e não há replicação do procedimento, como ocorreu neste trabalho (Engels & Diehr, 2003). Buscou-se adaptar as técnicas existentes para se considerar o cenário de repetição, a exemplo do que Zhou e colaboradores (2001) utilizam.

Para se verificar a eficiência de cada um dos valores imputados quando comparados aos valores verdadeiros, as medidas de *performance* aplicadas por Junger (2008) e Engels e Diehr (2003) poderiam ter sido utilizados. Os resultados desta aplicação seriam capazes de avaliar o procedimento de imputação em si, isto é, a variabilidade existente entre os coeficientes de cada banco de dados imputado, e não o valor combinado após a aplicação das regras de Rubin. A aplicação destas medidas não foi adotada neste trabalho por questões operacionais. Além disto, não se encontrou, até o presente momento, medidas de *performance* direcionadas à avaliação de imputação de variáveis categóricas, objeto de estudo neste trabalho.

Apesar dos resultados consistentes encontrados, este trabalho teve como base de comparação um banco de dados que se considerava completo, mas que, inicialmente, já apresentava um pequeno percentual (5,5%) de dados faltantes, que foram removidos para se obter um banco de dados 'padrão-ouro' para comparações. Explorações do banco de dados original (que incluía dados faltantes nas variáveis de interesse) demonstraram que o mecanismo de ocorrência poderia ser assumido como MAR e, a exclusão deste pequeno percentual de indivíduos parece não ter comprometido as estimativas obtidas com a imputação múltipla.

Finalmente, com este estudo, corrobora-se os achados de outros autores (van Buuren *et al.*, 1999; White & Royston, 2009), que, aplicando a imputação múltipla em análises de sobrevivência por meio do procedimento MICE, também encontraram resultados satisfatórios. A imputação múltipla via MICE pode ser aplicada quando da ocorrência de dados faltantes em estudos deste tipo, com variáveis de diversos tipos, incluindo categóricas, a fim de que as estimativas a serem obtidas nas análises sejam as mais fidedignas possíveis àquelas que seriam encontradas caso a não-resposta não ocorresse.

Conclusão

A imputação múltipla de dados parece ser uma maneira eficiente de se tratar a não-resposta em estudos de sobrevivência. A incorporação do tempo nos modelos de imputação pode se dar de forma tradicional, sem exigir transformações (logarítmicas, por exemplo) e o procedimento MICE parece fornecer resultados consistentes quando da aplicação na imputação de variáveis categóricas.

Referências bibliográficas

Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? International Journal of Methods in Psychiatric Research 2011; 20:40-49.

Boclin KdLS. Influência da posição sócio-econômica ao longo da vida nas desigualdades de cor/raça na ocorrência de miomas uterinos: Estudo Pró-Saúde. [Tese de Doutorado] Rio de Janeiro: Universidade do Estado do Rio de Janeiro; 2011.

Bono C, Ried L, Kimberlin C, Vogel B. Missing data on the Center for Epidemiologic Studies Depression Scale: A comparison of 4 imputation techniques. Research in Social and Administrative Pharmacy 2007; 3:1-27.

Canizares M, Barroso I, Alfonso K. [Methods for handling incomplete data in health research: a critical look]. Gac Sanit 2004; 18:58-63.

Catellier DJ, Hannan PJ, Murray DM, Addy CL, Conway TL, Yang S, et al. Imputation of missing data when measuring physical activity by accelerometry. Med Sci Sports Exerc 2005; 37:S555-62.

Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods 2001; 6:330-51.

Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006; 59:1087-91.

Enders CK. Applied Missing Data Analysis. 1st ed. New York: Guilford Press; 2010.

Engels JM, Diehr P. Imputation of missing longitudinal data: a comparison of methods. J Clin Epidemiol 2003; 56:968-76.

Faerstein E, Chor D, Lopes CdS, Werneck GL. Estudo Pró-Saúde: características gerais e aspectos metodológicos. Revista Brasileira de Epidemiologia 2005; 8:454-466.

Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995; 142:1255-64.

Haukoos JS, Newgard CD. Advanced Statistics: Missing Data in Clinical Research--Part 1: An Introduction and Conceptual Framework. Academic Emergency Medicine 2007; 14:662-668.

He Y. Missing data analysis using multiple imputation: getting to the heart of the matter. Circ Cardiovasc Qual Outcomes 2010; 3:98-105.

Junger WL. Análise, imputação de dados e interfaces computacionais em estudos de séries temporais epidemiológicas [Tese de Doutorado] Rio de Janeiro: Universidade do Estado do Rio de Janeiro; 2008.

Klebanoff MA, Cole SR. Use of Multiple Imputation in the Epidemiologic Literature. American Journal of Epidemiology 2008; 168:355-357.

Little RJA. Regression with missing X's: a review. Journal of the American Statistical Association 1992; 87:1227 - 1237.

Little RJA, Rubin DB. Statistical Analysis with missing data. 1st ed. New York: Wiley; 1987.

Moons KGM, Donders RART, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. Journal of Clinical Epidemiology 2006; 59:1092-1101.

Nunes LN, Kluck MM, Fachel JM. [Multiple imputations for missing data: a simulation with epidemiological data]. Cad Saude Publica 2009; 25:268-78.

Nunes LN, Klück MM, Fachel JMG. Comparação de métodos de imputação única e múltipla usando como exemplo um modelo de risco para mortalidade cirúrgica. Rev Bras Epidemiol 2010; 13:596-606.

R Development Core Team. R: A language and environment for statistical computing. Version 2.15. R Foundation for Statistical Computing, Vienna, Austria, 2012. http://www.r-project.org.

Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd ed. Piladelphia (PA): Lippincott Williams & Wilkins; 2008.

Rubin DB. Inference and Missing Data. Biometrika 1976; 63:581-592.

Rubin DB. Multiple Imputation for Nonresponse in Surveys. 1st. ed. New York: Wiley; 1987.

Schafer JL, Graham JW. Missing data: Our view of the state of the art. Psychological Methods 2002; 7:147-177.

Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009; 338:b2393-b2393.

Therneau TM. Survival package - A Package for Survival Analysis in S (R package), 2012.

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

Van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation 2006; 76:1049-1064.

Van Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software 2011; 45(3): 1-67.

White IR, Royston P. Imputing missing covariate values for the Cox model. Stat Med 2009; 28:1982-98.

White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine 2011; 30:377-399.

Zhou X-H, Eckert GJ, Tierney WM. Multiple imputation in public health research. Statistics in Medicine 2001; 20:1541-1549.

5.2 **Artigo 2**

"Imputação múltipla de dados – tutorial para aplicação em estudos epidemiológicos"

Introdução

A ocorrência de dados faltantes é um problema comum em estudos epidemiológicos (Rothman *et al.*, 2008), embora nem sempre este fato seja documentado quando da divulgação de resultados dos trabalhos. Ignorar a não-resposta pode ser potencialmente problemático, já que a ocorrência de dados faltantes em um pequeno percentual de algumas variáveis pode resultar em um grande número de observações com alguma informação não disponível, o que compromete a inferência sobre os parâmetros de interesse (Horton & Kleinman, 2007).

O problema vem atraindo atenção desde o fim da década de 70 (Rubin, 1976; Rubin, 1978), porém, no Brasil, pouca importância vem sendo dada ao tema. Considerando este cenário, o Estudo Pró-Saúde (EPS) - uma coorte entre funcionários técnico-administrativos de uma universidade no Rio de Janeiro, existente desde 1998 - tem demonstrado interesse em métodos estatísticos para se lidar com dados faltantes desde a década passada (Moreno, 2004). A vontade de se utilizar tais métodos advém da necessidade de se tratar adequadamente a não-resposta, que, a exemplo de outros estudos nacionais ou internacionais, também ocorre no EPS.

No Brasil, ainda são poucos os programas de investigação que resolvem lidar com o problema e que se dedicam ao uso de métodos avançados de análise para gerar estimativas válidas quando da ocorrência de dados faltantes.

A literatura sobre o tema apresenta algumas técnicas disponíveis para se tratar a não-resposta (Little & Rubin, 1987). O método mais aplicado, opção padrão dos programas estatísticos, é a chamada análise de observações completas, na qual apenas as observações com a totalidade de informação nas variáveis a serem estudadas são considerados na análise (Sinharay *et al.,* 2001). Em outras palavras, ocorre exclusão de todos os indivíduos que tenham dados faltantes em qualquer variável. Se um participante do estudo deixa de responder a alguma questão, toda a sua informação disponível será excluída. Este método pode enviesar as estimativas obtidas, e seu uso é desaconselhável (Schafer & Graham, 2002).

Outros métodos de análise de dados também utilizadas neste campo são as técnicas *ad hoc* e, dentre elas, cita-se: análise de observações disponíveis, análise de observações completas com ponderação, *inverse probability weighting* (IPW), imputação única (imputação pela média ou mediana, imputação por regressão, por regressão estocástica, imputação via *hot* ou *cold deck*, último valor observado); e técnicas que envolvem máxima verossimilhança, tais como o algoritmo EM. Little e Rubin (1987) e Allison (2001) e Enders (2010) trazem uma revisão sobre estes métodos, incluindo as vantagens, desvantagens e aplicações de cada um.

Não existe uma solução ideal para se lidar com qualquer problema de dados faltantes. Alguns autores (Harel & Zhou, 2007; Buhi *et al.,* 2008) afirmam que mesmo um pequeno percentual de não-resposta deve ser avaliado com cautela e, além disso, o pesquisador deve escolher o método que seja capaz de maximizar a acurácia e precisão das estimativas.

Nos últimos anos, maior atenção vem sendo dada à técnica de imputação múltipla, por ser considerada a mais adequada para se lidar com diversas situações de não-resposta (Klebanoff & Cole, 2008), porém, em recente revisão realizada nas bases de dados nacionais (LILACS e SCIELO), utilizando como palavras-chave "dados faltantes" e "imputação (múltipla)", e em bases de dados internacionais (PubMed/MedLine), utilizando as palavras "*missing data*", "*(multiple) imputation*" e "*Brazil*", apenas três publicações que tratam da questão dos dados faltantes aplicados à Epidemiologia foram encontradas no país (Nunes *et al.,* 2009; Nunes *et al.,* 2010; Camargos *et al.,* 2011).

Considerando a imputação múltipla a técnica apropriada a ser utilizada na prática dos estudos epidemiológicos, este trabalho pretende aprofundar os avanços realizados no Estudo Pró-Saúde nesta área, criando uma ferramenta que facilite o uso da técnica não só neste estudo, mas que seja passível de uso por qualquer grupo de pesquisa cujos dados tenham estrutura semelhante aos do EPS. Espera-se contribuir para divulgação da técnica ainda incipiente e pouco explorada no Brasil, especificamente na área de análise de sobrevivência, na qual o uso da imputação parece não estar totalmente consolidado inclusive no cenário internacional (White & Royston, 2009).

Tendo em vista a necessidade de textos metodológicos mencionada por outros autores brasileiros (Nunes *et al.,* 2010), este trabalho visa fornecer subsídios e colaborar para divulgação e uso eficiente da imputação múltipla de dados faltantes em pesquisas epidemiológicas.

Métodos

Imputação múltipla – considerações gerais

A técnica de IM foi proposta por Rubin na década de 1970 (Rubin, 1976) e possui características bayesianas, já que não se deseja imputar um valor único, mas uma distribuição preditiva dos valores faltantes, dado aquilo que foi observado. Basicamente, consiste em criar múltiplas cópias do banco de dados (‘*m*’ bancos, sendo $m>1$), com os dados faltantes substituídos por valores imputados. Para gerar tais valores, são usados os dados observados. Sequencialmente, podem ser utilizadas técnicas estatísticas tradicionais e o modelo de interesse é aplicado a cada banco gerado. Para combinação das estimativas, as “regras de Rubin” (Rubin, 1987) são então aplicadas, considerando a variabilidade dos resultados nos diferentes bancos imputados. Desta forma, a incerteza associada à imputação é levada em conta (Canizares *et al.*, 2004; Sterne *et al.*, 2009).

Os passos necessários para a imputação podem ser descritos da seguinte maneira: 1) especificação de um modelo paramétrico para os dados completos sob MAR; 2) suposição de uma distribuição *a priori* (não informativa) para os parâmetros desconhecidos do modelo; e, 3) simulação de múltiplas amostras independentes da distribuição *a posteriori* condicional dos valores faltantes, dada a informação observada (Enders, 2010; He, 2010). Em geral, os modelos de imputação seguirão este formato. Ao final, o último passo consiste em analisar e combinar as estimativas dos *m* bancos de dados gerados.

A IM possui como vantagens a possibilidade de se usar métodos de análise para bancos completos e a habilidade de incluir o conhecimento do pesquisador (Nunes, 2007). E, apesar de Rubin (1987) mencionar que o procedimento pode ser mais exaustivo quando comparado a outras técnicas de imputação, com o aumento da capacidade dos computadores, o esforço necessário à IM passou a ser menor. Ainda sim, alguns passos importantes são necessários para correta aplicação da técnica. Maiores detalhes serão expostos nas seções seguintes.

Passo 1 – Avaliar pressupostos necessários à imputação

Para que a execução da imputação múltipla seja adequada, o primeiro passo é a adoção de alguns pressupostos quanto ao mecanismo de ocorrência dos dados faltantes e quanto ao padrão desta. Ao se avaliar a pertinência da adoção de tais pressupostos, pode-se decidir se a imputação é indicada para lidar com o problema da não-resposta em um estudo particular.

Quanto aos mecanismos, estes podem ser: dado faltante completamente aleatório (em inglês, *missing completely at random* – MCAR), dado faltante aleatório (*missing at random* – MAR) e dado faltante não aleatório (*missing not at random* – MNAR).

Dados faltantes completamente aleatórios (MCAR) ocorrem quando a informação faltante é independente daquilo que foi observado e do que não o foi. Neste caso, os dados observados constituem uma subamostra aleatória dos dados totais (Little & Rubin, 1987) – em geral, esse pressuposto não costuma ser a realidade dos estudos epidemiológicos (Grittner *et al.*, 2011).

Já o dado faltante aleatório (MAR) ocorre quando a informação faltante depende apenas do que foi observado, e, tal informação pode ser reconstituída com base nos dados disponíveis. Como afirmam Grittner e colaboradores (2011), o dado faltante pode diferir entre os subgrupos de uma amostra, mas é aleatório dentro de cada um deles. Tanto MAR quanto MCAR são considerados "dados faltantes ignoráveis": quando ocorrem, pode-se ignorar as razões que explicam porque o dado não foi observado e empregar uma técnica apropriada para lidar com o problema (Buhi *et al.*, 2008). Com base no pressuposto de os dados faltantes serem do tipo MAR, grande parte das técnicas de imputação, incluindo a IM, foi desenvolvida (Rubin, 1987; Donders *et al.*, 2006).

Dado faltante não aleatório (MNAR) acontece quando a informação faltante depende do que não foi observado e, possivelmente, daquilo que foi de fato observado. É também chamado "dado faltante não ignorável" (Rubin, 1987), e os modelos de análise comumente utilizados não são capazes de lidar adequadamente com os dados incompletos resultantes deste mecanismo (Buhi *et al.*, 2008). Segundo Nunes (2007), dados mais sujeitos a serem MNAR são aqueles encontrados nos extremos da distribuição, com valores maiores ou menores do que o padrão observado na amostra.

Um exemplo hipotético destes mecanismos no âmbito do Estudo Pró-Saúde diz respeito à pergunta sobre realização de exame de mama pelas mulheres, em um estudo que se

dedicasse a avaliar saúde da mulher. Caso uma participante deixe de responder por ter faltado ao trabalho, ou estar de férias ou de licença, o dado faltante poderia ser completamente aleatório (MCAR). Poderia ser não aleatório (MNAR) caso a mulher deixe de responder a pergunta exatamente por não realizar o exame rotineiramente. E, caso a resposta permanecesse em branco por razões não relacionadas à não realização do exame, mas a outras questões presentes no questionário (por exemplo, à renda ou à escolaridade), o dado faltante poderia ser considerado aleatório (MAR).

Ainda em relação aos mecanismos de ocorrência de dados faltantes, determinar se este é do tipo MAR ou MNAR depende dos dados e da plausibilidade em se assumir um mecanismo ou outro, já que a adoção deste pressuposto não é testável (Little & Rubin, 1987). Diversos autores afirmam que, caso o dado faltante seja inevitável em determinado estudo, pode-se fazer com que ele se aproxime o máximo possível do mecanismo MAR e se afaste de MNAR através da inclusão de medidas adicionais durante o planejamento do estudo, o que faz com que a probabilidade do dado faltante ser MAR aumente (Harel & Zhou, 2007; Buhi *et al.*, 2008).

Quanto ao padrão de ocorrência dos dados faltantes, de maneira geral, este pode ser monotônico ou não-monotônico. No primeiro, comum em estudos longitudinais, blocos de variáveis apresentam cada vez mais dados faltantes, ou de acordo com as ondas de seguimento ou de acordo com a sequência de variáveis em uma mesma onda. No segundo, tal padrão crescente de não-resposta não é observado (Little & Rubin, 1987). Há métodos específicos para lidar com um ou outro (Haukoos & Newgard, 2007), e os modelos utilizados para padrões monotônicos são considerados mais simples (Horton & Kleinman, 2007).

Cabe ressaltar que a não-resposta pode ocorrer em uma unidade inteira – quando o sujeito selecionado se recusa a continuar no estudo – ou em alguns itens – nos quais os indivíduos se negam a responder algumas perguntas por constrangimento, por exemplo. Neste sentido, a IM pode ser utilizada quando a não-resposta ocorre em um ou mais itens de um indivíduo e não na unidade toda – ou seja, requer que os dados daquele indivíduo com algum dado faltante possuam alguma porção de informação observada (Haukoos & Newgard, 2007).

Tendo considerado o mecanismo de ocorrência do dado faltante e seu padrão, mantendo a decisão pela imputação, pode-se seguir para a próxima etapa.

Passo 2 – Tomar decisões para aplicação da IM

Algumas decisões importantes devem ser tomadas antes da aplicação da técnica. Um aspecto fundamental e complexo quando da realização de IM é a especificação do modelo de imputação, que se subdivide em duas outras decisões: a forma do modelo (se linear, logístico, polinomial, etc.) e o conjunto de variáveis preditoras que o irão compor. Van Buuren e colaboradores (1999) afirmam que a função principal do modelo de imputação é gerar uma amplitude de valores plausíveis para aquelas variáveis sendo imputadas. A forma do modelo e seus parâmetros não são de grande interesse, o que torna a escolha exata da forma funcional do modelo de IM de menor importância. Portanto, a seleção de variáveis se torna, juntamente com a quantidade de bancos de dados a ser gerada com a imputação, uma decisão essencial para que o método possa ser corretamente aplicado. A seguir, estes dois passos serão descritos.

Passo 2A - Selecionar variáveis que irão compor o modelo de imputação

Quanto à seleção de co-variáveis do modelo de IM, este deve ser capaz de preservar todas as associações importantes entre as variáveis (Patrician, 2002) e, deve incluir: as variáveis que farão parte do modelo de análise a ser aplicado posteriormente (incluindo o desfecho) (Moons *et al.,* 2006); variáveis auxiliares – que podem não fazer parte do modelo de análise, mas possuem alguma relação com os dados faltantes; e, variáveis que fazem parte do desenho do estudo (Haukoos & Newgard, 2007).

Alguns autores (van Buuren *et al.,* 1999; Spratt *et al.,* 2010; Camargos *et al.,* 2011) afirmam que uma estratégia de seleção inclusiva, que abarque a maior quantidade possível de variáveis, usando a maior quantidade disponível possível de informação, levará a um viés menor e precisão máxima. Outros autores (Allison, 2001; Patrician, 2002; Newgard & Haukoos, 2007) também mencionam que, na adição de co-variáveis, é melhor errar incluindo mais variáveis do que o necessário do que o contrário. Collins e colaboradores (2001), ao compararem estratégias inclusivas (com o maior número possível de variáveis) e restritivas (com o menor número possível), encontraram melhores resultados nos procedimentos de IM com abordagem inclusiva.

A ideia de se adicionar a maior quantidade de co-variáveis ao modelo tem relação com o fato de que, ao se incluir tantos preditores quanto possível, pode-se garantir mais facilmente de que o mecanismo dos dados faltantes seja MAR, reduzindo a necessidade de ajustes para considerar a possibilidade de MNAR (Schafer, 1997; van Buuren *et al.,* 1999). Entretanto, em um banco de dados com centenas de variáveis, em um estudo de maior porte com diferentes blocos de questões que tratam de diferentes temas, como é o caso do Estudo Pró-Saúde, a adição de todas as variáveis disponíveis não é confiável e necessária - por conta de multicolinearidade e de problemas computacionais.

A literatura disponível sobre IM não fornece uma regra geral para guiar esta seleção de preditores, o que faz com que alguns autores optem por executar um procedimento *stepwise* para selecionar estas variáveis (Howard *et al.,* 2011). O que se afirma é que variáveis com distribuições altamente assimétricas podem fazer parte do modelo transformadas, para que se aproximem da normalidade (Patrician, 2002). As variáveis devem entrar de forma completa – isto é, categorizações devem ser feitas posteriormente, para que não haja restrição dos valores plausíveis a serem imputados (Newgard & Haukoos, 2007). O mesmo pode ser estendido para a imputação de escores – pode ser mais apropriado imputar as variáveis originais e depois reconstruir o escore com base nelas (Azur *et al.,* 2011). Além disto, a inclusão do desfecho de interesse do modelo de análise posterior parece reduzir o viés na imputação das suas co-variáveis (Moons *et al.,* 2006) e reter a associação entre o desfecho e estes preditores (que serão os desfechos do modelo de imputação) (He, 2010). Outro aspecto consensual é que variáveis candidatas a preditoras no modelo de IM não devem conter um alto percentual de dados faltantes (Nunes, 2007).

Uma abordagem para seleção de variáveis é sugerida por van Buuren e colaboradores (1999) e consiste em quatro passos:

1. Incluir todas as variáveis que aparecem no modelo de análise – se isto não for feito, haverá viés na análise posterior dos bancos imputados, principalmente se o modelo de análise contiver relações de predição fortes.
2. Incluir as variáveis relacionadas à (não) resposta – fatores que influenciam a ocorrência do dado faltante e outras variáveis de interesse cujas distribuições variem entre os grupos de respondentes e não-respondentes. Pode-se encontrar essas variáveis através do cálculo da correlação entre

elas com uma indicadora (*dummy*) da não resposta naquela variável de interesse.

3. Adicionar variáveis que explicam uma porção considerável da variância da variável de interesse – preditores como esses ajudam a reduzir a incerteza das imputações. Também podem ser identificadas por meio da análise de correlação.
4. Por fim, remover dos conjuntos determinados nos passos 3 e 4 variáveis que tenham um alto percentual de dados faltantes dentro do subgrupo de casos incompletos.

Todos esses aspectos a serem considerados fazem com que a seleção de variáveis que irão compor o modelo de imputação se torne um passo importante, que deve ser cuidadosamente pensado e que considere não somente o modelo teórico-conceitual que subjaz o objeto de estudo, mas também o modelo capaz de predizer a ocorrência dos dados faltantes.

Passo 2B – Considerar a quantidade de bancos de dados (*m*) a ser criada com a imputação

Quanto à decisão sobre a quantidade de banco de dados (*m*) a ser gerado, apesar de muitos autores (Kristman *et al.*, 2005; Harel & Zhou, 2007) utilizarem valores entre 3-10, Graham e colaboradores (2007) afirmam que o *m* necessário pode ser muito maior do que aquele que vem sendo utilizado em algumas análises, e que se deve levar em conta o percentual de informação faltante – que difere do percentual de dados faltantes em situações mais complexas – e a eficiência que se deseja alcançar com a imputação, seguindo a equação proposta por Rubin (1987). Azur e colaboradores (2011) comentam que, na prática, gerar um *m* elevado pode ser problemático e demandar muito tempo, mas, também afirmam que a decisão sobre a quantidade deve se basear no tamanho do banco de dados, na quantidade de informação não observada e nos recursos computacionais disponíveis. De maneira geral, para se alcançar uma eficiência relativa de 100% com o procedimento de imputação, seria necessário $m \to \infty$ e um *m* elevado levaria a uma situação computacionalmente intensiva.

Passo 3 – Executar a técnica de IM

Uma última importante decisão diz respeito à forma como a IM será conduzida. He (2010) propõe uma classificação dos modelos para realização da imputação múltipla em duas grandes categorias: *Joint Modelling* (JM) e *Sequential Regression Multiple Imputation* – ou, como mencionado por outros autores – *Fully Conditional Specification* (FCS) (van Buuren *et al.*, 2006).

A abordagem classificada como *Joint Modelling* (JM) divide as observações em grupos com características de não-resposta semelhantes e imputa os dados faltantes com cada padrão de acordo com um modelo conjunto para todas as variáveis que sejam comuns a todas as observações (He, 2010). Em outras palavras, JM envolve a especificação de uma distribuição multivariada para os dados faltantes e, na sequência, a amostragem de imputações da distribuição *a posteriori* condicional desses dados através de cadeias de Markov-Monte Carlo (MCMC), por exemplo. Entre os exemplos clássicos estão o uso de modelos normais multivariados para variáveis contínuas, modelos log-lineares para variáveis categóricas e modelos de efeitos mistos para medidas repetidas ou análises multinível (Schafer, 1997). O método de JM é atrativo caso uma distribuição multivariada forneça uma descrição razoável dos dados (van Buuren & Groothuis-Oudshoorn, 2011), mas pode ser problemático por possuir pouca flexibilidade, necessária para representar estruturas de dados complexas que são observadas em muitos estudos (He, 2010). Harel & Zhou (2007) trazem uma breve revisão de alguns métodos classificados como JM.

Para lidar com determinadas estruturas de dados de maior complexidade, pode-se optar pelos métodos da classe FCS, cuja característica principal é a especificação de modelos condicionais separados para cada variável incompleta, utilizando as demais como preditoras. A cada passo do algoritmo da FCS, as imputações são geradas para os valores faltantes de uma variável, e, esses valores imputados são usados na imputação da próxima variável, e o processo se repete até atingir a convergência (He, 2010). FCS pode ser mais adequada do que os métodos de JM quando não há uma distribuição multivariada apropriada para lidar com o problema que se tem e quando maior flexibilidade é necessária quando da construção dos modelos (van Buuren *et al.*, 2006). Um problema da FCS é o fato de que, em algumas situações mais complexas, o modelo pode não convergir, tornando a aplicação do método inviável (van Buuren *et al.*, 2006; Enders, 2010).

O procedimento de FCS mais recentemente utilizado é o de Imputação Múltipla por equações em cadeia – MICE (*Multiple Imputation by Chained Equations*), que consiste em:

> 1. Uma imputação única (como imputação por regressão ou pela média) é realizada em cada valor faltante no banco de dados. Esse valor imputado pode ser considerado como um "marcador de posição".
> 2. Os "marcadores de posição" de uma das variáveis – a primeira na sequência dos modelos determinados - (VAR 1) são removidos, e seus valores voltam a ser dados faltantes.
> 3. Os valores observados de "VAR 1" do passo 2 são regredidos nas outras variáveis do modelo de IM, que pode ou não consistir de todas as variáveis no conjunto de dados. "VAR 1" seria a variável dependente no modelo de regressão e todas as demais são as co-variáveis deste modelo.
> 4. Os valores faltantes de "VAR 1" são então substituídos pelas predições (imputações) do modelo de regressão – são substituídos por amostras da distribuição preditiva *a posteriori* desta variável. Esta amostragem ocorre utilizando o amostrador de Gibbs. Na sequência, quando "VAR 1" é utilizada como variável independente nos modelos de imputação de outras variáveis (VAR 2, VAR 3, ... , VAR *K*), tanto os dados observados quanto essas imputações são utilizadas (Azur *et al.*, 2011).

Os passos 2-4 são repetidos para cada variável que tenha dado faltante. O ciclo de cada variável constitui uma iteração. No fim de um ciclo, todos os valores faltantes são substituídos com as predições oriundas das regressões que refletem as relações existentes na porção observada dos dados. As imputações são atualizadas e retidas ao fim de cada ciclo, o que gera um banco de dados imputado. A quantidade de ciclos pode ser determinada pelo pesquisador e, em geral, a convergência será alcançada quando as estimativas dos parâmetros de interesse estiverem estáveis (White *et al.*, 2011).

A característica mais importante do MICE é sua habilidade em lidar com diferentes tipos de variáveis, cada uma com seu próprio modelo de imputação, que é atualizado com base no modelo anterior, e utilizado para determinar o modelo posterior. Zhou e colaboradores (2001) também afirmam que o procedimento evita que ocorra extrapolação dos limites (amplitude) dos dados e, o fato de se usar amostras da distribuição preditiva *a*

posteriori ajuda a preservar distribuições e associações entre as variáveis. A maior complexidade do procedimento reside na necessidade de se especificar modelos de imputação distintos para cada variável, diferentemente da abordagem de JM, na qual um único modelo de imputação é especificado (He, 2010; Azur *et al.*, 2011). Ainda sim, considera-se o MICE a técnica mais flexível e adequada na realidade dos estudos epidemiológicos e, neste trabalho, esta será a técnica a ser exemplificada.

Passo 4 – Combinar os resultados obtidos

Após tomar as decisões necessárias e já tendo criado os bancos de dados imputados com a técnica selecionada, a análise tradicional é realizada separadamente em cada um deles, gerando uma quantidade de estimativas pontuais tão grande quanto a quantidade de bancos de dados criados (*m*). A partir daí, a combinação dessas estimativas para gerar um valor único se dá a partir das chamadas "regras de Rubin" (Rubin, 1987). Para determinação da estimativa pontual média (o coeficiente da análise de regressão, por exemplo), determina-se uma média aritmética simples dividindo o somatório dos valores dos coeficientes encontrados em cada banco pela quantidade (*m*) de bancos gerados na imputação.

Para se determinar a variância combinada ($\bar{T}$) das estimativas, considera-se a variância dentro das imputações ($\bar{U}_m$) e a variância entre elas (B_m), segundo as equações propostas por Rubin (1987). Além disto, para amostras razoavelmente grandes, o intervalo de 95% de confiança para será dado segundo a aproximação assintótica. As estimativas do risco relativo (ou do *hazard ratio*) podem ser determinadas com base no exponencial do valor determinado como estimativa pontual média dos *m* bancos e os limites dos intervalos de confiança podem ser calculados da maneira tradicional, também por aproximação assintótica (van Buuren *et al.*, 1999). As equações podem ser encontradas com detalhes em Rubin (1987) e van Buuren e colaboradores (1999).

Os aplicativos estatísticos capazes de executar a IM costumam disponibilizar a opção de combinação dos resultados das análises dos *m* bancos de dados segundo estas regras.

Caso particular - IM em estudos de sobrevivência

Na Epidemiologia, poucos são os trabalhos dedicados a lidar com a questão da IM em estudos de sobrevivência de maneira metodológica, independente do modelo de análise utilizado (paramétrico, semi-paramétrico ou não paramétrico). No que tange à seleção das variáveis a compor o modelo e ao procedimento geral de imputação, a técnica ocorre de maneira semelhante às demais e parece fornecer resultados satisfatórios ao se aplicar MICE (van Buuren *et al.*, 1999; White & Royston, 2009).

A particularidade quando da construção dos modelos de imputação em sobrevivência diz respeito à inclusão do desfecho: nos modelos de regressão, deve-se considerar o status (censura ou caso) e o tempo de contribuição de cada participante ao estudo como desfechos, e não só o status do indivíduo. Alguns autores (van Buuren *et al.*, 1999; White & Royston, 2009) discutem de que maneira o tempo deve fazer parte dos modelos de imputação e afirmam que esta inclusão pode se dar de diversas formas: adicionando o status, o tempo e o logaritmo do tempo como preditores no modelo de imputação; adicionando apenas o status e o logaritmo do tempo; e outras. Acredita-se que melhores resultados podem ser obtidos se o tempo for acrescentado ao modelo de imputação (Ali *et al.*, 2011), porém, ainda não há consenso sobre a forma desta inclusão (White & Royston, 2009).

Desta forma, além das considerações quanto aos pressupostos necessários à adoção de IM, às variáveis que o irão compor, à quantidade de bancos de dados a ser gerado (*m*), e, quanto à maneira que a imputação ocorrerá – se via JM ou FCS, a forma como o tempo de sobrevida será incorporado ao modelo constitui a principal distinção da IM em modelos de sobrevivência, exigindo atenção especial do analista.

Aplicação

Para exemplificar a aplicação da técnica de IM em um cenário particular da análise de sobrevivência, selecionou-se uma das configurações propostas por (*Artigo 1*), embasadas nas análises anteriormente conduzidas por Boclin (2011). Utilizando dados do Estudo Pró-Saúde, ilustrar-se-á, a seguir, a aplicação da IM quando da ocorrência de 10% de dados faltantes no mecanismo MAR.

Estudo Pró-Saúde e análise de referência

O Estudo Pró-Saúde é um estudo longitudinal realizado entre funcionários técnico-administrativos de uma universidade localizada no Rio de Janeiro, cujo início se deu em 1998. Entre seus objetivos principais figuram a investigação do papel de determinantes sociais no estado de saúde dos indivíduos (Faerstein *et al.*, 2005).

Encontra-se atualmente em sua quarta etapa, sendo as anteriores realizadas nos anos de 1999 (Fase 1), 2001 (Fase 2) e 2006 (Fase 3). Para a realização deste trabalho, foram usadas as informações provenientes da linha de base (Fases 1 e 2) somente. Detalhes sobre a população de estudo encontram-se em Faerstein e colaboradores (2005). O EPS foi aprovado pelo Comitê de Ética em Pesquisa da instituição na qual é conduzido.

Selecionou-se como base para exemplificação a análise realizada por Boclin (2011), em tese de doutorado recentemente defendida. Uma das propostas de tal trabalho foi averiguar se posição sócio-econômica durante a infância, início da vida adulta, ou ao longo da vida era mediadora da relação entre cor/raça e a ocorrência de miomas uterinos (MU) na população. Para conduzir tal investigação, utilizou-se os dados da linha de base do EPS para a população feminina, que contava com 1819 participantes, dentre 2466 funcionárias elegíveis (73,8% das elegíveis foram avaliadas em 1999-2001).

O desfecho utilizado na análise foi o diagnóstico médico auto-relatado de miomas uterinos e outras informações coletadas acerca dos miomas uterinos foram a idade da participante quando do diagnóstico do mioma e sobre a realização de cirurgia de retirada do útero (histerectomia) e a idade da mulher quando submetida a esta, para que se reconstituísse adequadamente o período de seguimento das participantes.

A cor/raça das participantes, utilizada como variável de exposição principal,foi coletada através da pergunta aberta e agrupada da seguinte maneira: branca, parda, preta e amarela. A categoria amarela foi excluída das análises por apresentar um número pequeno de participantes (n=8, 0,44%), que poderiam introduzir ‘ruídos’ às análises.

Outras variáveis utilizadas no modelo como marcadoras de acesso e utilização de serviços de saúde foram: Plano de saúde; Realização de teste Papanicolaou; Realização de exame de mama. A variável “plano de saúde” foi coletada em três categorias: 1 - Sim, como titular; 2 - Sim, como dependente; 3 – Não, e, posteriormente, dicotomizada. As variáveis “realização de teste Papanicolaou” e “realização de exame de mama” foram coletadas em

quatro categorias e posteriormente dicotomizadas em ‘nunca realizou ou realizou há mais de 3 anos’ e ‘realizou há menos de 3 anos’.

A escolaridade da participante também fazia parte do modelo em questão, tendo sido coletada em sete categorias e, posteriormente, transformada em três: 0 - até 1º grau completo; 1 – até 2º grau completo e 2 – universitário completo ou mais. Outra variável importante foi a idade da participante, inserida na forma de tercis.

Para lidar com o problema de os dados terem sido coletados transversalmente (dados da linha de base do EPS), buscou-se resgatar as histórias de seguimento das participantes, a partir das informações por elas relatadas. Assumiu-se que o seguimento se iniciou aos 20 anos de idade para todas as mulheres e o tempo final de “acompanhamento” delas foi determinado da seguinte maneira: para mulheres que não apresentavam miomas uterinos, utilizou-se a idade em 1999; para as mulheres que relataram miomas, a idade do diagnóstico do mesmo, e, para mulheres que relataram histerectomia, a idade da cirurgia. Determinou-se como tempo final de seguimento máximo a idade de 50 anos. Desta forma, dez casos de miomas diagnosticados após esse período foram censurados.

Nas análises multivariadas, foram utilizados modelos de riscos proporcionais de Cox para estimar a Razão de Hazards com intervalos de 95% de confiança (IC 95%) (Boclin, 2011). O modelo utilizado neste trabalho é o que tem por desfecho o status da participante (caso - 1; censura – 0) e o tempo de seguimento conforme os critérios antes mencionados. Como variáveis de exposição estão a cor/raça e a idade das participantes e como fatores de confusão do modelo, as variáveis marcadoras de acesso aos serviços de saúde (plano de saúde, exame de Papanicolaou, exame de mama) e a escolaridade da mulher.

Utilizando o mesmo conjunto de dados analisado por Boclin (2011), procedeu-se a exclusão das participantes consideradas inelegíveis para o estudo de base e das mulheres de cor/raça amarela.

Com as exclusões necessárias, o banco de dados inicial, com 1819 mulheres, passou a ser constituído de 1593 participantes. Destas, 5,5 % (n=88) não tinham informação em alguma das variáveis do modelo proposto ou outras variáveis a serem utilizadas durante os procedimentos de simulação ou imputação. Com a remoção destas últimas, o banco de dados considerado completo consistiu de 1505 mulheres, todas com informações completas no desfecho, idade, cor/raça, plano de saúde, realização de exame de mama e Papanicolaou, escolaridade e cor/raça de acordo com a classificação proposta pelo IBGE.

Este banco de dados foi organizado de forma a permitir uma análise utilizando o modelo de Cox. Em seguida, a partir do banco de dados considerado completo, simulou-se a ocorrência de dados faltantes na variável cor/raça (*Artigo 1*) para posterior imputação e comparação dos resultados obtidos com aqueles do banco de dados original (n=1505).

Para permitir livre acesso aos procedimentos adotados, todas as etapas de montagem do banco de dados, simulação e imputação foram conduzidas utilizando o *software* R (R Development Core Team, 2012), versão 2.15. A decisão pelo R se deu principalmente por se tratar de um *software* livre, de fácil acesso por qualquer pesquisador, e que cobre de maneira satisfatória as análises necessárias tanto na simulação (conduzida para fins ilustrativos neste estudo) e na IM quanto na análise de sobrevivência.

O principal pacote do R para realizar imputações por equações em cadeia – técnica a ser ilustrada aqui - é o "MICE", criado por van Buuren e Groothuis-Oudshoorn (2011). O "MICE" permite a aplicação da técnica em diferentes cenários, com diferentes tipos de variáveis (normais, binárias, multinomiais, etc), bem como a combinação das estimativas geradas (pelas "regras de Rubin"), diagnóstico dos modelos de imputação, alterações na matriz de predição a ser utilizada para IM, além da determinação de quais variáveis do banco de dados serão imputadas e de que maneira o serão. Também inclui a possibilidade de análise de sensibilidade para se considerar um mecanismo de dados faltantes do tipo MNAR. O R conta ainda com outros pacotes para realização de imputação: Amelia, mitools, mix, pan, dentre outros.

Horton e Kleinman (2007) trazem uma revisão sobre os programas disponíveis para realização de imputação múltipla. Atualmente, todos os grandes *softwares* estatísticos dispõem de ferramentas para execução da técnica.

Imputação múltipla de dados

Para esta análise, após simulação da ocorrência de 10% de dados faltantes do tipo MAR na variável cor/raça *(Artigo 1)*, procedeu-se à imputação múltipla, considerando no modelo de IM as variáveis: idade (em tercis), escolaridade (três categorias), plano de saúde (dicotomizada), realização de exame de mama, realização de exame de Papanicolaou (também

dicotomizadas), status da participante (se caso ou censura) e tempo de contribuição no estudo (em anos). Optou-se por não transformar o tempo e incluí-lo diretamente no modelo.

Antes de se incluir as variáveis no modelo de imputação, considerou-se quais informações disponíveis no banco de dados do EPS poderiam estar relacionadas à não-resposta na variável cor/raça. Decidiu-se pelas variáveis idade e escolaridade, que, por já fazerem parte do modelo de análise estabelecido a priori, já seriam incluídas nos modelos de IM. Portanto, o modelo de imputação aqui especificado inclui as mesmas variáveis do modelo de análise.

Por se tratar de uma análise de sobrevivência, a seleção pelo MICE como técnica de imputação pareceu adequada. Como a variável a ser imputada era multinomial, a opção 'polyreg' do pacote MICE (van Buuren & Groothuis-Oudshoorn, 2011) foi selecionada. Quanto à quantidade de banco de dados a ser gerada, as opções m=5, m=10, m=20 e m=100 foram testadas.

Para estas especificações, os seguintes comandos foram necessários:

```
require (mice)   # Carregar a biblioteca necessária
ini <- mice(banco_miss, maxit=0, pri=F)    # Criar as imputações - para m=5
pred <- ini$pred # Tirar variáveis da matriz de predição e não imputar as que não interessam
pred[,c("id","idadeind","stringpadrao")] <- 0      # Retirar da predição essas variáveis e deixar apenas as de interesse.
imp <- mice(banco_miss,m=5, meth=c("","","","","","","polyreg","","","","","",""), pred=pred)  # Selecionar as variáveis que serão imputadas (neste caso, apenas uma) e o método de imputação dela – 'polyreg', bem como a quantidade de bancos de dados gerados (m=5)
```

Maiores detalhes quanto às alterações necessárias na matriz de predição das variáveis e quanto à seleção de variáveis que devem ou não ser imputadas, bem como outras opções disponíveis na biblioteca MICE podem ser encontradas em van Buuren e Groothuis-Oudshoorn (2011). Vale ressaltar que a quantidade de bancos de dados (*m*) a ser gerada pode ser alterada, sendo o *default* da função mice m=5.

Sequencialmente, cada banco de dados gerado (neste caso, m=5) foi analisado, de acordo com o modelo estabelecido *a priori* no estudo de base, utilizando o modelo de riscos proporcionais de Cox da mesma maneira: como desfecho utilizou-se o status e o tempo; como

variáveis de exposição a cor/raça das participantes já imputada e como co-variáveis a idade em tercis, plano de saúde, realização de exame de mama e Papanicolaou (dicotomizadas) e a escolaridade da participante (em três categorias). Para evitar problemas de convergência no modelo de Cox, o número máximo de iterações foi aumentado para 100. Ao final das análises, os valores de cada banco de dados foram combinados, obedecendo as "regras de Rubin" anteriormente expostas.

Para análise e combinação dos resultados dos 'm' bancos, utilizou-se os comandos:

```
require(survival)  # Carregar a biblioteca necessária para a análise de sobrevivência
# Analisar cada banco completo com o modelo de analise estabelecido a priori:
y<-Surv(banco_miss$tempo,banco_miss$status)
fit <- with(imp,coxph(y~corecat + as.factor(tercildeidade) + as.factor(m7) + as.factor(m8) + as.factor(b1) + as.factor(escolaridade), data=banco_miss, control = coxph.control (iter.max=100))
a <- summary(pool(fit))  # combinação dos resultados dos bancos de dados imputados pelas 'regras de Rubin'.
```

No estudo original de simulação, cujo propósito era determinar a eficiência da IM e viabilidade de aplicação em estudos que utilizam análises de sobrevivência (*Artigo 1*), desejava-se determinar a distribuição das estimativas obtidas em cada cenário diferente de imputação, e, para isto, o processo foi replicado 100 vezes e, ao final, foi determinada a média Monte Carlo dos coeficientes do modelo de regressão para a variável de interesse (cor/raça) e seus erros-padrão, bem como a variância entre os coeficientes.

Utilizando as médias dos coeficientes e erros, determinou-se a Razão de Hazards e seu intervalo de confiança (IC95%) em cada cenário, a fim de compará-los com os valores do padrão-ouro e com os da análise de observações completas. Estas últimas também foram replicadas 100 vezes, a fim de se determinar a distribuição das estimativas desta análise, após simulação dos dados faltantes, da mesma forma como ocorreu com a IM. Ao final das replicações, os procedimentos foram avaliados através de indicadores de *performance*. Detalhes quanto aos cenários de simulação, imputação e avaliação podem ser encontrados no *Artigo 1*.

Resultados

A análise do banco de dados considerado completo (‘padrão-ouro’, n=1505) forneceu as seguintes estimativas de Razão de Hazards (e respectivos intervalos de confiança de 95%): para mulheres de cor/raça branca – RH=1,0; para pardas, RH= 1,1365 (0,8534 – 1,5137); e, para pretas, RH= 1,7925 (1,3805 – 2,3276). Tais valores são compatíveis com os valores de RH encontrados no mesmo modelo de análise do estudo de base (Boclin, 2011). Vale ressaltar que, neste estudo de referência, o procedimento adotado foi o de análise de observações completas.

Neste mesmo cenário (MAR, 10% de dados faltantes), a análise de observações completas, isto é, a análise do banco de dados original excluindo 10% de dados (n=1354), forneceu como resultados os valores: RH = 1,1269 (0,8311 – 1,5281) para pardas e 1,7841 (1,3520 – 2,3543) para pretas. As estimativas pontuais são semelhantes àquelas do banco de dados original, porém, a amplitude dos intervalos de confiança é superior aos valores originais, por conta da diminuição do número de indivíduos analisados. Apesar de pequena – por conta da manutenção do ‘n’ elevado, mesmo com 10% de não-resposta, ocorreu perda de poder estatístico, aumentando a amplitude dos intervalos de confiança para as duas categorias.

Utilizando os comandos anteriormente expostos, a imputação múltipla foi aplicada à variável cor/raça, gerando as estimativas de RH apresentadas na tabela 1. Os diferentes *m* utilizados (m=5, 10, 20 e 100) apontam que, para este percentual de dados faltantes, m=10 seria capaz de produzir o resultado mais próximo do valor ‘verdadeiro’, tanto para pardas quanto para pretas. Demais *m* também fornecem estimativas semelhantes às originais, com uma discrepância maior para pardas quando m=100 e com m=20 para pretas.

A análise de *performance* desta ilustração também apresenta resultados condizentes com o esperado (tabela 2). As medidas de Desvio Médio Quadrático (DMQ) e Viés indicam que as estimativas obtidas com a imputação (e replicação do procedimento) são semelhantes aos valores de referência. A variância, calculada entre os coeficientes estimados ao final de cada replicação, demonstra pequena dispersão entre estes valores, indicando a qualidade das simulações, imputações e replicações.

Tabela I – Resultados da imputação múltipla nas estimativas da Razão de Hazards para a variável cor/raça, após simulação de 10% de dados faltantes no mecanismo MAR. Dados do Estudo Pró-Saúde-RJ (1999-2001).

	Pardas	Pretas
m=5	1,1325 (0,8351 – 1,5359)	1,8028 (1,3662 – 2,3788)
m=10	1,1348 (0,8373 – 1,5379)	1,7905 (1,3568 – 2,3627)
m=20	1,1384 (0,8397 – 1,5433)	1,8117 (1,3736 – 2,3894)
m=100	1,1266 (0,8311 – 1,5272)	1,8000 (1,3645 – 2,3746)

Tabela II – Avaliação da *performance* do procedimento de simulação e imputação de dados no cenário MAR, 10%. Dados do Estudo Pró-Saúde-RJ (1999-2001).

		DMQ	Viés	Variância
m=5	Pardas	0,0047	0,0035	0,0022
	Pretas	0,0044	-0,0057	0,0019
m=10	Pardas	0,0056	0,0015	0,0032
	Pretas	0,0052	0,0011	0,0028
m=20	Pardas	0,0051	-0,0016	0,0027
	Pretas	0,0046	-0,0106	0,0020
m=100	Pardas	0,0052	0,0088	0,0027
	Pretas	0,0044	-0,0042	0,0019

Conclusão

A imputação múltipla, se bem conduzida, pode ser uma ferramenta útil para se lidar com a questão da não-resposta, inclusive em estudos de sobrevivência. Os passos necessários

para sua aplicação devem ser cuidadosamente planejados e a reflexão quanto à melhor forma de se imputar deve ser realizada em cada situação particular.

No caso aqui apresentado, a imputação foi capaz de fornecer resultados satisfatórios quando o mecanismo de ocorrência era do tipo MAR e o percentual de dados faltantes era de 10%, porém, sabe-se que a imputação pode ser efetiva em outros cenários (*Artigo 1*).

Em relação à inclusão do tempo de contribuição ao estudo nos modelos de imputação, incorporá-lo da maneira tradicional (em anos) parece suficiente, e, a adição do desfecho neste modelo foi fundamental para reter a associação entre a variável de exposição principal imputada e o desfecho no modelo de análise, fornecendo resultados semelhantes àqueles do estudo de referência (Boclin, 2011).

No que diz respeito às replicações apresentadas neste estudo, vale ressaltar que, apesar de se mostrarem mais eficientes quando da utilização da imputação (*Artigo 1*), estas não precisam ser utilizadas quando da aplicação da imputação múltipla em situações 'reais' na epidemiologia.

Quanto às medidas de *performance* aqui apresentadas, novamente, estas foram determinadas para se verificar a capacidade da imputação múltipla em produzir estimativas pontuais e intervalos de confiança não viesados e consistentes e, de maneira geral, também não precisariam ser determinadas em situações nas quais a imputação múltipla não seja objeto de estudo, mas sim uma ferramenta estatística fundamental para se obter medidas de associação, coeficientes, taxas, e outras estimativas de maneira confiável quando da ocorrência de dados faltantes.

Por fim, com este trabalho, espera-se que a técnica de imputação múltipla passe a ser empregada de forma adequada nos estudos epidemiológicos (de sobrevivência ou não), para que o problema da não-resposta possa ser (finalmente) considerado e analisado de maneira adequada.

Referências bibliográficas

Ali AM, Dawson SJ, Blows FM, Provenzano E, Ellis IO, Baglietto L, et al. Comparison of methods for handling missing data on immunohistochemical markers in survival analysis of breast cancer. Br J Cancer 2011; 104:693-9.

Allison PD. Missing Data. 1st ed. Thousand Oaks, CA: Sage, 2001.

Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? International Journal of Methods in Psychiatric Research 2011; 20:40-49.

Boclin KdLS. Influência da posição sócio-econômica ao longo da vida nas desigualdades de cor/raça na ocorrência de miomas uterinos: Estudo Pró-Saúde. [Tese de Doutorado] Rio de Janeiro: Universidade do Estado do Rio de Janeiro; 2011.

Buhi ER, Goodson P, Neilands TB. Out of sight, not out of mind: strategies for handling missing data. Am J Health Behav 2008; 32:83-92.

Camargos VP, Cesar CC, Caiaffa WT, Xavier CC, Proietti FA. [Multiple imputation and complete case analysis in logistic regression models: a practical assessment of the impact of incomplete covariate data]. Cad Saude Publica 2011; 27:2299-313.

Canizares M, Barroso I, Alfonso K. [Methods for handling incomplete data in health research: a critical look]. Gac Sanit 2004; 18:58-63.

Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods 2001; 6:330-51.

Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006; 59:1087-91.

Enders CK. Applied Missing Data Analysis. 1st ed. New York: Guilford Press; 2010.

Faerstein E, Chor D, Lopes CdS, Werneck GL. Estudo Pró-Saúde: características gerais e aspectos metodológicos. Revista Brasileira de Epidemiologia 2005; 8:454-466.

Graham JW, Olchowski AE, Gilreath TD. How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. Prevention Science 2007; 8:206-213.

Grittner U, Gmel G, Ripatti S, Bloomfield K, Wicki M. Missing value imputation in longitudinal measures of alcohol consumption. Int J Methods Psychiatr Res 2011; 20:50-61.

Harel O, Zhou X-H. Multiple imputation: review of theory, implementation and software. Statistics in Medicine 2007; 26:3057-3077.

Haukoos JS, Newgard CD. Advanced Statistics: Missing Data in Clinical Research--Part 1: An Introduction and Conceptual Framework. Academic Emergency Medicine 2007; 14:662-668.

He Y. Missing data analysis using multiple imputation: getting to the heart of the matter. Circ Cardiovasc Qual Outcomes 2010; 3:98-105.

Horton NJ, Kleinman KP. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. Am Stat 2007; 61:79-90.

Howard G, McClure LA, Moy CS, Safford MM, Cushman M, Judd SE, et al. Imputation of Incident Events in Longitudinal Cohort Studies. American Journal of Epidemiology 2011; 174:718-726.

Klebanoff MA, Cole SR. Use of Multiple Imputation in the Epidemiologic Literature. American Journal of Epidemiology 2008; 168:355-357.

Kristman VL, Manno M, Côté P. Methods to Account for Attrition in Longitudinal Data: Do They Work? A Simulation Study. European Journal of Epidemiology 2005; 20:657-662.

Little RJA, Rubin DB. Statistical Analysis with missing data. 1st ed. New York: Wiley, 1987.

Moons KGM, Donders RART, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. Journal of Clinical Epidemiology 2006; 59:1092-1101.

Moreno AB. Mobilidade ocupacional e qualidade de vida entre funcionários de uma universidade no Rio de Janeiro: o Estudo Pró-Saúde [Tese de Doutorado] Rio de Janeiro: Universidade do Estado do Rio de Janeiro; 2004.

Newgard CD, Haukoos JS. Advanced statistics: missing data in clinical research--part 2: multiple imputation. Acad Emerg Med 2007; 14:669-78.

Nunes LN. Métodos de imputação de dados aplicados na área da saúde [Tese de Doutorado] Porto Alegre: Universidade Federal do Rio Grande do Sul; 2007.

Nunes LN, Kluck MM, Fachel JM. [Multiple imputations for missing data: a simulation with epidemiological data]. Cad Saude Publica 2009; 25:268-78.

Nunes LN, Klück MM, Fachel JMG. Comparação de métodos de imputação única e múltipla usando como exemplo um modelo de risco para mortalidade cirúrgica. Rev Bras Epidemiol 2010; 13:596-606.

Patrician PA. Multiple imputation for missing data. Res Nurs Health 2002; 25:76-84.

R Development Core Team. R: A language and environment for statistical computing. Version 2.15. R Foundation for Statistical Computing, Vienna, Austria, 2012. http://www.r-project.org.

Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd ed. Philadelphia (PA): Lippincott Williams & Wilkins; 2008.

Rubin DB. Inference and Missing Data. Biometrika 1976; 63 (3):581-592.

Rubin DB. Multiple imputation in sample surveys – a phenomenological bayesian approach to nonresponse. In: Proceedings of the Survey Research Methods Section, American Statistical Association 1978.

Rubin DB. Multiple Imputation for Nonresponse in Surveys. 1st ed. New York: Wiley; 1987.

Schafer JL. Analysis of incomplete multivariate data. 1st ed. Boca Raton, Florida: Chapman & Hall/CRC; 1997.

Schafer JL, Graham JW. Missing data: Our view of the state of the art. Psychological Methods 2002; 7:147-177.

Sinharay S, Stern HS, Russell D. The use of multiple imputation for the analysis of missing data. Psychol Methods 2001; 6:317-29.

Spratt M, Carpenter J, Sterne JAC, Carlin JB, Heron J, Henderson J, et al. Strategies for Multiple Imputation in Longitudinal Studies. American Journal of Epidemiology 2010; 172:478-487.

Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009; 338:b2393-b2393.

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

Van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation 2006; 76:1049-1064.

Van Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software 2011; 45 (3): 1-67.

White IR, Royston P. Imputing missing covariate values for the Cox model. Stat Med 2009; 28:1982-98.

White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine 2011; 30:377-399.

Zhou X-H, Eckert GJ, Tierney WM. Multiple imputation in public health research. Statistics in Medicine 2001; 20:1541-1549.

CONSIDERAÇÕES FINAIS

Nesta dissertação, evidenciou-se a utilidade da imputação múltipla como ferramenta para se lidar com o problema dos dados faltantes. Em todas as situações apresentadas, a imputação e replicação dos procedimentos produziram resultados semelhantes aos encontrados no banco de dados 'completo', mesmo quando o cenário não era rigorosamente mais indicado para a aplicação da técnica.

Vale ressaltar que o banco de dados considerado 'padrão-ouro' nesta dissertação também apresentava um pequeno percentual de dados faltantes inicialmente, mas acredita-se que tal questão não invalide os resultados e as comparações aqui conduzidas. Na realidade dos estudos epidemiológicos, sobretudo os longitudinais, encontrar um conjunto de dados completo, sem dados faltantes, é uma tarefa árdua, senão impossível. Além disto, um dos objetivos aqui propostos era a aplicação da imputação múltipla na realidade epidemiológica e não em conjuntos de dados simulados, que, possivelmente não refletiriam as estruturas dos dados coletados na prática.

Pretende-se que o tutorial construído nesta dissertação sirva como ferramenta para as análises realizadas no Estudo Pró-Saúde nas quais os dados faltantes constituam um problema. Além disto, a IM pode ser aplicada por qualquer grupo de pesquisa cujos dados se assemelhem aos aqui apresentados e, o material produzido nesta dissertação poderá embasar a utilização da IM em outros cenários, não só em análises de sobrevivência.

Quanto à esta, espera-se ter ajudado a preencher a lacuna identificada quanto à incorporação do tempo nos modelos de sobrevivência, e ressalta-se a importância de que esta decisão seja tomada após a avaliação do contexto no qual a análise está inserida.

Para o Estudo Pró-Saúde, o tutorial terá aplicação imediata, tanto para os dados da fase quatro do estudo, quanto para variáveis de outras fases, que apresentam percentuais consideráveis de dados faltantes e que necessitem de imputação para que sejam utilizadas adequadamente nas análises.

Finalmente, imputar os dados, sejam eles do tipo MCAR, MAR e MNAR, de maneira consciente, adotando os pressupostos e tomando as decisões pertinentes, é mais apropriado do que ignorar a não-resposta e sequer mencioná-la quando das análises de dados. Menciona-se que a imputação deve ser considerada particularmente em cada análise, incorporando as relações e peculiaridades de cada uma.

REFERÊNCIAS

Ali AM, Dawson SJ, Blows FM, Provenzano E, Ellis IO, Baglietto L, et al. Comparison of methods for handling missing data on immunohistochemical markers in survival analysis of breast cancer. Br J Cancer 2011; 104:693-9.

Allison PD. Missing Data. 1st ed. Thousand Oaks, CA: Sage, 2001.

Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? International Journal of Methods in Psychiatric Research 2011; 20:40-49.

Baneshi MR, Talei AR. Multiple imputation in survival models: applied on breast cancer data. Iran Red Crescent Med J 2011; 13:544-9.

Boclin KdLS. Influência da posição sócio-econômica ao longo da vida nas desigualdades de cor/raça na ocorrência de miomas uterinos: Estudo Pró-Saúde. [Tese de Doutorado] Rio de Janeiro: Universidade do Estado do Rio de Janeiro; 2011.

Bono C, Ried L, Kimberlin C, Vogel B. Missing data on the Center for Epidemiologic Studies Depression Scale: A comparison of 4 imputation techniques. Research in Social and Administrative Pharmacy 2007; 3:1-27.

Buhi ER, Goodson P, Neilands TB. Out of sight, not out of mind: strategies for handling missing data. Am J Health Behav 2008; 32:83-92.

Burns RA, Butterworth P, Kiely KM, Bielak AAM, Luszcz MA, Mitchell P, et al. Multiple imputation was an efficient method for harmonizing the Mini-Mental State Examination with missing item-level data. Journal of Clinical Epidemiology 2011; 64:787-793.

Camargos VP, Cesar CC, Caiaffa WT, Xavier CC, Proietti FA. [Multiple imputation and complete case analysis in logistic regression models: a practical assessment of the impact of incomplete covariate data]. Cad Saude Publica 2011; 27:2299-313.

Canizares M, Barroso I, Alfonso K. [Methods for handling incomplete data in health research: a critical look]. Gac Sanit 2004; 18:58-63.

Carvalho MS, Andreozzi WL, Codeço CT, Campos DP, Barbosa MTS, Shimakura SE. Análise de Sobrevivência - Teoria em aplicações em saúde. 2ª edição. Rio de Janeiro: Editora Fiocruz; 2011.

Catellier DJ, Hannan PJ, Murray DM, Addy CL, Conway TL, Yang S, et al. Imputation of missing data when measuring physical activity by accelerometry. Med Sci Sports Exerc 2005; 37:S555-62.

Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods 2001; 6:330-51.

Cox DR. Regression Models on Life-tables. Journal of the Royal Statistical Society 1972; Series B, 34:187-220.

Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006; 59:1087-91.

Enders CK. Applied Missing Data Analysis. 1st ed. New York: Guilford Press; 2010.

Engels JM, Diehr P. Imputation of missing longitudinal data: a comparison of methods. J Clin Epidemiol 2003; 56:968-76.

Faerstein E, Chor D, Lopes CdS, Werneck GL. Estudo Pró-Saúde: características gerais e aspectos metodológicos. Revista Brasileira de Epidemiologia 2005; 8:454-466.

Graham JW, Olchowski AE, Gilreath TD. How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. Prevention Science 2007; 8:206-213.

Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995; 142:1255-64.

Grittner U, Gmel G, Ripatti S, Bloomfield K, Wicki M. Missing value imputation in longitudinal measures of alcohol consumption. Int J Methods Psychiatr Res 2011; 20:50-61.

Harel O, Zhou X-H. Multiple imputation: review of theory, implementation and software. Statistics in Medicine 2007; 26:3057-3077.

Haukoos JS, Newgard CD. Advanced Statistics: Missing Data in Clinical Research--Part 1: An Introduction and Conceptual Framework. Academic Emergency Medicine 2007; 14:662-668.

He Y. Missing data analysis using multiple imputation: getting to the heart of the matter. Circ Cardiovasc Qual Outcomes 2010; 3:98-105.

Horton NJ, Kleinman KP. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. Am Stat 2007; 61:79-90.

Hosmer DW, Lemeshow S, May S. Applied Survival Analysis: Regression Modeling of Time to Event Data. 2nd ed. New York: Wiley, 2011.

Howard G, McClure LA, Moy CS, Safford MM, Cushman M, Judd SE, et al. Imputation of Incident Events in Longitudinal Cohort Studies. American Journal of Epidemiology 2011; 174:718-726.

Junger WL. Análise, imputação de dados e interfaces computacionais em estudos de séries temporais epidemiológicas [Tese de Doutorado] Rio de Janeiro: Universidade do Estado do Rio de Janeiro; 2008.

Klebanoff MA, Cole SR. Use of Multiple Imputation in the Epidemiologic Literature. American Journal of Epidemiology 2008; 168:355-357.

Kmetic A, Joseph L, Berger C, Tenenhouse A. Multiple imputation to account for missing data in a survey: estimating the prevalence of osteoporosis. Epidemiology 2002; 13:437-44.

Kristman VL, Manno M, Côté P. Methods to Account for Attrition in Longitudinal Data: Do They Work? A Simulation Study. European Journal of Epidemiology 2005; 20:657-662.

Little RJA. Regression with missing X's: a review. Journal of the American Statistical Association 1992; 87:1227 - 1237.

Little RJA, Rubin DB. Statistical Analysis with missing data. 1st ed. New York: Wiley, 1987.

Mishra GD, Dobson AJ. Multiple imputation for body mass index: lessons from the Australian Longitudinal Study on Women's Health. Statistics in Medicine 2004; 23:3077-3087.

Moons KGM, Donders RART, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. Journal of Clinical Epidemiology 2006; 59:1092-1101.

Moreno AB. Mobilidade ocupacional e qualidade de vida entre funcionários de uma universidade no Rio de Janeiro: o Estudo Pró-Saúde [Tese de Doutorado] Rio de Janeiro: Universidade do Estado do Rio de Janeiro; 2004.

Mumford SL, Schisterman EF, Gaskins AJ, Pollack AZ, Perkins NJ, Whitcomb BW, et al. Realignment and multiple imputation of longitudinal data: an application to menstrual cycle data. Paediatr Perinat Epidemiol 2011; 25:448-59.

Newgard CD, Haukoos JS. Advanced statistics: missing data in clinical research--part 2: multiple imputation. Acad Emerg Med 2007; 14:669-78.

Nunes LN. Métodos de imputação de dados aplicados na área da saúde [Tese de Doutorado] Porto Alegre: Universidade Federal do Rio Grande do Sul; 2007.

Nunes LN, Klück MM, Fachel JM. [Multiple imputations for missing data: a simulation with epidemiological data]. Cad Saude Publica 2009; 25:268-78.

Nunes LN, Klück MM, Fachel JMG. Comparação de métodos de imputação única e múltipla usando como exemplo um modelo de risco para mortalidade cirúrgica. Rev Bras Epidemiol 2010; 13:596-606.

Paik MC. Multiple imputation for the Cox proportional hazards model with missing covariates. Lifetime Data Anal 1997; 3:289-98.

Patrician PA. Multiple imputation for missing data. Res Nurs Health 2002; 25:76-84.

R Development Core Team. R: A language and environment for statistical computing. Version 2.15. R Foundation for Statistical Computing, Vienna, Austria, 2012. http://www.r-project.org.

Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd ed. Philadelphia (PA): Lippincott Williams & Wilkins; 2008.

Rubin DB. Inference and Missing Data. Biometrika 1976; 63(3):581-592.

Rubin DB. Multiple imputation in sample surveys – a phenomenological bayesian approach to nonresponse. In: Proceedings of the Survey Research Methods Section, American Statistical Association 1978.

Rubin DB. Multiple Imputation for Nonresponse in Surveys. 1st ed. New York: Wiley; 1987.

Schafer JL. Analysis of incomplete multivariate data. 1st ed. Boca Raton, Florida: Chapman & Hall/CRC; 1997.

Schafer JL. Multiple imputation: a primer. Stat Methods Med Res 1999; 8:3-15.

Schafer JL, Graham JW. Missing data: Our view of the state of the art. Psychological Methods 2002; 7:147-177.

Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res 2011; 0:1-18.

Seaman SR, White IR, Copas AJ, Li L. Combining multiple imputation and inverse-probability weighting. Biometrics 2012; 68:129-37.

Sinharay S, Stern HS, Russell D. The use of multiple imputation for the analysis of missing data. Psychol Methods 2001; 6:317-29.

Spratt M, Carpenter J, Sterne JAC, Carlin JB, Heron J, Henderson J, et al. Strategies for Multiple Imputation in Longitudinal Studies. American Journal of Epidemiology 2010; 172:478-487.

Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009; 338:b2393-b2393.

Therneau TM. Survival package - A Package for Survival Analysis in S (R package), 2012.

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

Van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation 2006; 76:1049-1064.

Van Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software 2011; 45 (3): 1-67.

Wang CN, Little R, Nan B, Harlow SD. A hot-deck multiple imputation procedure for gaps in longitudinal recurrent event histories. Biometrics 2011; 67:1573-82.

White IR, Royston P. Imputing missing covariate values for the Cox model. Stat Med 2009; 28:1982-98.

White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine 2011; 30:377-399.

Zhou X-H, Eckert GJ, Tierney WM. Multiple imputation in public health research. Statistics in Medicine 2001; 20:1541-1549.

APÊNDICE A – Descrição geral das variáveis utilizadas neste trabalho

Em caráter informativo, são apresentadas as frequências absolutas, relativas e intervalos de confiança gerados a partir do banco de dados completo (n=1505),

	n	% (IC95%)
Cor/raça		
Brancas	796	52,9% (50,3 – 55,4)
Pardas	352	23,4% (21,3- 25,6)
Pretas	357	23,7% (21,6 – 25,9)
Ocorrência de miomas uterinos (status)		
Casos	337	22,4% (20,3 – 24,6)
Censuras	1168	77,6% (75,4 – 80,0)
Tercis de idade		
1º tercil (36)	568	37,7% (35,3 – 40,2)
2º tercil (43)	491	32,6% (30,2 – 35,1)
3º tercil	446	29,6% (27,3 – 32,0)
Plano de Saúde		
Possuem	960	63,8% (61,3 – 66,2)
Não possuem	545	36,2% (33,8 – 38,8)
Realização de exame de mama		
Nunca realizaram/ Realizaram há mais de 3 anos	171	11,4% (9,8 – 13,1)
Realizaram há menos de 3 anos	1334	88,6% (87,0 – 90,2)
Realização de exame de Papanicolaou		
Nunca realizaram/	171	11,4% (9,8 – 13,1)

Realizaram há mais de 3 anos		
Realizaram há menos de 3 anos	1334	88,6% (87,0 – 90,2)
Escolaridade		
Até 1º grau completo	260	17,3% (15,4 – 19,3)
2º grau completo	529	35,1% (32,7 – 37,6)
Universitário completo ou mais	716	47,6% (45,0 – 50,1)

APÊNDICE B – Scripts utilizados na elaboração desta dissertação

Montagem do banco com dados faltantes – remoção de inelegiveis de acordo com a figura 2.

```
setwd("C:\\Documents and Settings\\Owner\\My Documents\\Dropbox\\Thais\\Análises Dissertação p. Washington\\Limpeza e montagem do banco full")
banco1=read.csv2('prosaudelinhadebase_1819.csv')


# Exclusão 1: quem não respondeu a M14:
banco2 = subset(banco1,(!is.na(m14)))
dim(banco2)   # 1733 -  86 exclusões

# Exclusao 2: quem respondeu 1 em M14 e não respondeu m15:
banco3 = subset(banco2,(!(m14==1&(is.na(m15)))))
dim(banco3)  # 1726 - 7 exclusões

# Exclusão 3: quem não respondeu a m18:
banco4 = subset(banco3,(!is.na(m18)))
dim(banco4)  # 1609 - 117 exclusões

# Exclusão 4: quem respondeu 1 em m18 e não respondeu m19:
banco5 = subset(banco4,(!(m18==1&(is.na(m19)))))
dim(banco5)  # 1606 - 3 exclusões

# Exclusão 5: mulheres com m14=1 e m15<20
banco6 = subset(banco5,((m14==1&m15>=20)|(m14==2&is.na(m15))))
dim(banco6)  # 1603 - 3 exclusões

# Exclusão 6: mulheres com m18=1 e m19<20
banco7 = subset(banco6,((m18==1&m19>=20)|(m18==2&is.na(m19))))
dim(banco7)  # 1602 - 1 exclusão

# Exclusão 7: mulheres de cor/raça amarela (corecat=4)
banco8 = subset(banco7,((is.na(corecat))|(corecat==1)|(corecat==2)|(corecat==3)))
dim(banco8)  # 1593 - 9 exclusões


# Salvando o banco:
write.csv2(banco8,file="analise_1593.csv",row.names=F,na="")
```

Montagem do banco de dados para análise de sobrevivência

```
# Banco a ser utilizado:  analise_1593

setwd("C:\\Documents and Settings\\Owner\\My Documents\\Dropbox\\Thais\\Análises Dissertação p. Washington\\Limpeza e montagem do banco full")
dados <-read.csv2("analise_1593.csv")
attach(dados)
require(epicalc)

# Preparo do banco para o modelo de COX (incluindo tempo e status)

## Recodificação do desfecho - Diagnóstico médico de mioma uterino auto-relatado.
#Nome da variável nas análises= mio
#Categorias=  não=0, sim=1
tab1(m14)
mio<-ifelse(m14==2,0,ifelse(m14==1,1,NA)) ## recodificando a variável para O/1
tab1(mio)
dados$mioma <- mio


# Criando a variável de status:
status <- ifelse(mio==1&m15<=50,1,0)
tab1(status)     # mulheres com mioma até os 50 entraram - só considerei censura após 50.
dados$status <- status      # Tivemos 10 censuras (m14=1 e status=0)

# Criando a variável tempo
tempo.ini <- 20
dados$tempo.ini <- tempo.ini

write.csv2(dados, file="analise_1593_COX.csv",na="",row.names=F)
```

Montagem do banco de dados completo (“Padrão-ouro”), sem dados faltantes nas variáveis de interesse.

```
setwd("C:\\Documents and Settings\\Owner\\My Documents\\Dropbox\\Thais\\Análises Dissertação p. Washington\\Limpeza e montagem do banco full")
banco <-read.csv2("analise_1593_COX.csv")
require(epicalc)
attach(banco)


# Variáveis do modelo escolhido - panorama geral:

tab1(status)  # 0 missing
```

```
tab1(tempo)   # 0 missing
tab1(idadeind) # 0 missing
tab1(corecat)  # 3.2% de missing
tab1(b1)  # 0.5% de missing
tab1(m7)  # 0.5% de missing
tab1(m8)   # 0.5% de missing
tab1(e35)  # 1.1% de missing
tab1(e39) # 0.5% de missing

# Excluindo o NA das variáveis do modelo que contém missing:
banco2 =  banco [complete.cases (banco$b1, banco$m7, banco$m8, banco$corecat ,banco$e35, banco$e39 ), ]
dim(banco2) # Tem 5,5% de missing em relação a 1593.

# Salvando o banco com a escolaridade limpa:
write.csv2(banco2,file="analise_esc_racaibgefull.csv",row.names=F,na="")
```

Análise do banco de dados completo – "padrão-ouro" (n=1505)

```
# Leitura do banco:
setwd("C:\\Documents and Settings\\Owner\\My Documents\\Dropbox\\Thais\\Análises Dissertação p. Washington\\Cox - banco full")
full = read.csv2 ("analise_esc_racaibgefull.csv")
require(survival)
require(epicalc)
attach(full)

# Limpando e deixando apenas as variáveis de interesse:

full2=subset(full,
select=c("id","numquest","p2quest","b1","e35","e39","m7","m8","m14","idadeind","corecat", "mioma","status","tempo.ini","tempo.fim","tempo"))
detach(full)
attach(full2)

# Passo 1 - análise exploratória uni e bivariada:

# A) Construção das variáveis:

## A1) Desfecho
#Diagnóstico médico de mioma uterino auto-relatado.
#Nome da variável nas análises= mio
#Categorias=  não=0, sim=1
tab1(m14)
```

```
mio<-ifelse(m14==2,0,ifelse(m14==1,1,NA)) ## recodificando a variável para O/1
tab1(mio)

# Analisando o status - 1 - caso; 0 - censusa
tab1(status)  # 10 censuras

## A2) Exposição principal
## Cor/raça
#Nome da variável nas análises= raca
#Categorias=  branca=1, parda=2 e preta=3.
tab1(corecat)
raca = corecat

## A3) Outras variáveis:

##Idade - Idade em tercil
##Nome da variável nas análises= tercildeidade
# B) Recategorizando idade (tercis):
tercil = quantile(idadeind,probs=c(1/3,2/3))
tercil

idadeterc = ifelse(idadeind<=36, "tercil1", ifelse(idadeind>36&idadeind<=43, "tercil2",ifelse
(idadeind>43,"tercil3",NA)))
tab1(idadeterc)

## B1 - Plano de saúde
#Nome da variável nas análises= plano
#Categorias= sim=0, não= 1
tab1(b1)
plano<-ifelse(b1<3,0,ifelse(b1==3,1,NA)) # dicotomizando tem ou não plano.
tab1(plano)

# M7 - Realização de Papanicolaou
#Categorias= realizou há menos de 3 anos= 0, nunca realizou/realizou há mais de 3 anos= 1
tab1(m7)
papa<-ifelse(m7<3,1,ifelse(m7>=3,0,NA))
tab1(papa)

## M8 -  Realização de  exame de mama
#Categorias= realizou há menos de 3 anos= 0, nunca realizou/realizou há mais de 3 anos= 1
tab1(m8)
mama<-ifelse(m8<3,1,ifelse(m8>=3,0,NA))
tab1(mama)

# Recategorizando escolaridade:
#Nome da variavel nas analises= esc
#Categorias=  universitario completo= 2, 2º grau completo= 1,ate 1º grau completo= 0
```

```
tab1(e35)
esc<-ifelse(e35<4,0,ifelse(e35>=4&e35<6,1,ifelse(e35>=6,2,NA)))
tab1(esc)

# A4) Categorização de algumas variáveis:

raca<-factor(raca)
papa<-factor(papa)
mama<-factor(mama)
plano<-factor(plano)
esc <- as.factor(esc)


# Passo 2 - análise de sobrevida

#Modelo de Cox ajustado por idade e variáveis de acesso a serviços de saúde e escolaridade
reg1<-coxph(y~raca+idadeind+plano+mama+papa+esc,x=T)
summary(reg1)
```

Simulação – montagem dos bancos de dados

MCAR

Banco a ser utilizado: analise_1593_COX # n=1593 - já foram removidas as inelegiveis para o meu estudo, na ordem do fluxograma, mas ainda há missing nas covariáveis de interesse. Já pronto para o modelo de Cox.

A partir dele, o banco será limpo e organizado para simulação de MCAR, imputação e análise via modelo de COX.

```
setwd("C:\\Documents and Settings\\Owner\\My Documents\\Dropbox\\Thais\\Análises Dissertação p. Washington\\Simulação e imputação")
dados <-read.csv2("analise_1593_23-01_COX.csv")
attach(dados)
require(epicalc)


### Limpeza e recodificação de variáveis:

# A) Recategorizando escolaridade:
#Nome da variável nas análises= esc
#Categorias= universitário completo= 2, 2º grau completo= 1,até 1º grau completo= 0
tab1(e35)
```

```
esc<-ifelse(e35<4,0,ifelse(e35>=4&e35<6,1,ifelse(e35>=6,2,NA)))
tab1(esc)
dados$escolaridade = esc

# B) Recategorizando idade (tercis):
tercil = quantile(idadeind,probs=c(1/3,2/3))
tercil

idadeterc =
ifelse(idadeind<=36,"tercil1",ifelse(idadeind>36&idadeind<=43,"tercil2",ifelse(idadeind>43,
"tercil3",NA)))
tab1(idadeterc)
dados$tercildeidade = idadeterc

# C) Arrumando a variavel plano de saude (dicotomizando) # Categorias= sim=1, nao= 0
tab1(b1)
plano<-ifelse(b1<3,1,ifelse(b1==3,0,NA))
table(plano)
dados$b1=plano  # substituir no banco para facilitar as coisas

# D) Arrumando a variável M7 - Realização de Papanicolaou

#Categorias= realizou há menos de 3 anos= 0, nunca realizou/realizou há mais de 3 anos= 1
tab1(m7)
papa<-ifelse(m7<3,1,ifelse(m7>=3,0,NA))
dados$m7 = papa
tab1(dados$m7)

## E) Arrumando a variável M8-  Realização de  exame de mama
#Categorias= realizou há menos de 3 anos= 0, nunca realizou/realizou há mais de 3 anos= 1
tab1(m8)
mama<-ifelse(m8<3,1,ifelse(m8>=3,0,NA))
dados$m8 = mama
tab1(dados$m8)

# # Variáveis do modelo escolhido - panorama geral:

tab1(status)  # 0 missing
tab1(tempo)   # 0 missing
tab1(idadeind) # 0 missing
tab1(corecat)  # 3.2% de missing
tab1(dados$b1)  # 0.5% de missing
tab1(dados$m7)  # 0.5% de missing
tab1(dados$m8)   # 0.5% de missing
tab1(esc) # 1.1% de missing
tab1(e39)  #0.5% de missing
```

```
# Excluindo o NA das variáveis do modelo que contém missing (incluindo a pergunta sobre cor/raça do IBGE (E39):
banco <- dados [complete.cases (dados$b1, dados$m7, dados$m8, dados$corecat, dados$esc, dados$esc1, dados$esc2, dados$esc3, dados$e39),]
dim(banco) # Tem aproximadamente 5,5% de missing em relação a 1505.


## Salvando o banco de dados para análise:

banco2 = subset(banco, select= c("id", "numquest", "p2quest", "b1", corraca", "e35", "e39", "m7", "m8", "idadeind","corecat", "escolaridade", "tercildeidade", "mioma", "status", "tempo.ini", "tempo.fim", "tempo"))

write.csv2(banco2,file="analise_1505_MCAR.csv",row.names=F,na=" ")



MAR

# Banco a ser utilizado:  analise_1593_COX # n=1593

setwd("C:\\Documents and Settings\\Owner\\My Documents\\Dropbox\\Thais\\Análises Dissertação p. Washington\\Simulação e imputação")
dados <-read.csv2("analise_1593_23-01_COX.csv")
attach(dados)
require(epicalc)

# Avaliacao dos missings da variavel corecat:

# 1) Criando dummy de missing dela:

tab1(corecat)
miss = ifelse(!is.na(corecat),0,1) # 1 se for missing; 0 se n?o
tab1(Demissie, et al.)      # 51 missings

# Decidimos utilizar apenas as variaveis escolaridade e idade (em tercis)

# 2) Recategorizando escolaridade:
#Nome da variavel nas analises= esc
#Categorias=  universitario completo= 2, 2º grau completo= 1,ate 1º grau completo= 0
tab1(e35)
esc<-ifelse(e35<4,0,ifelse(e35>=4&e35<6,1,ifelse(e35>=6,2,NA)))
tab1(esc)
dados$escolaridade = esc
# 3) Recategorizando idade (tercis):
# Fazendo via ifelse:
```

```
tercil = quantile(idadeind,probs=c(1/3,2/3))
tercil

idadeterc =
ifelse(idadeind<=36,"tercil1",ifelse(idadeind>36&idadeind<=43,"tercil2",ifelse(idadeind>43,
"tercil3",NA)))
tab1(idadeterc)
dados$tercildeidade = idadeterc


### Avaliando escolaridade (esc) e idade (tercis) de acordo com ser missing ou nao:

#a) Fazendo as analises bivariadas:

## Escolaridade:
tabpct(esc,miss)   # Alguma tendencia de ser mais alto em niveis menores de escolaridade

## Idade:
tabpct(idadeterc,miss) # Aparentemente,  mais missing entre os mais velhos.


## Criando dummies para idade e escolaridade:

# IDADE:
idadeterc1=ifelse(idadeterc=="tercil1",1,0)
dados$idadeterc1 = idadeterc1

idadeterc2=ifelse(idadeterc=="tercil2",1,0)
dados$idadeterc2 = idadeterc2

idadeterc3=ifelse(idadeterc=="tercil3",1,0)
dados$idadeterc3 = idadeterc3


 # ESCOLARIDADE:

esc1 = ifelse(esc==0,1,0) # Ate 1º grau completo
dados$esc1 = esc1

esc2 = ifelse(esc==1,1,0) # Ate 2º grau completo
dados$esc2 = esc2

esc3 = ifelse(esc==2,1,0) # Univ completo ou mais
dados$esc3 = esc3
# MODELO MULTIVARIADO:
```

```
# Incluindo ESCOLARIDADE E IDADE EM TERCIS - DUMMIES E SEM A LINHA DE BASE:

mod1=glm(miss~ as.factor(esc2) + as.factor(esc3) + as.factor(idadeterc2) + as.factor(idadeterc3), family="binomial")
summary(mod1)

# Selecionando os padroes
banco_modelo=model.matrix(mod1)
padroes=unique(banco_modelo)

# Transformando em data frame para o apply rodar direito:
padroes = as.data.frame(padroes)

# Colapsando os padroes para facilitar:
padroes$stringpad=apply(padroes,1,paste,collapse="")
padroes

# Identificando os padroes:
id = seq(1,9,1)
padrao=cbind(id,padroes)
padrao

### Verificando probabilidade de missing em cada padrao:

# Antes - precisei rodar novamente esta parte pq %% não aceita dataframe
banco_modelo=model.matrix(mod1)
padroes=unique(banco_modelo)

# Agora sim, a prob de missing:
preditor=padroes %*% coef(mod1)
prob=exp(preditor)/(1+exp(preditor))

### Contando quantos individuos devem ser removidos de cada padrao, no banco com missing (n=1593)

npadrao=integer(nrow(padroes))
for(j in 1:nrow(padroes))
        #for(i in 1:nrow(banco_modelo))

        npadrao[j]=nrow(subset(banco_modelo,apply(banco_modelo,1,paste,collapse='')==paste(padroes[j,],collapse='')))

padrao=cbind(padrao,npadrao) # acrescentar a quantidade de pessoas que há em cada padrão

m_1593=npadrao*prob
```

```
#Arredondando m para cima:
m_1593=ceiling(m_1593)

miss_padrao= cbind(id,prob,m_1593)
miss_padrao  # Banco com os "m" baseados em 1593

# Criando um banco com os padroes separados, colapsados e a prob associada a cada um:
bancoprob=cbind(padrao,prob)
bancoprob  ## listagem de padrões existentes com os ids, strings e probabilidades de cada um.

### Fazendo os padrões no banco limpo (n= 1505)
## Baseando m em 1505 (n do banco final)

# Organizando algumas variáveis do banco de dados:

# A) Arrumando a variavel plano de saude (dicotomizando) # Categorias= sim=1, nao= 0
tab1(b1)
plano<-ifelse(b1<3,1,ifelse(b1==3,0,NA))
table(plano)
dados$b1=plano  # substituir no banco para facilitar as coisas

# B) Arrumando a variável M7 - Realização de Papanicolaou
#Categorias= realizou há menos de 3 anos= 0, nunca realizou/realizou há mais de 3 anos= 1
tab1(m7)
papa<-ifelse(m7<3,1,ifelse(m7>=3,0,NA))
dados$m7 = papa
tab1(dados$m7)

## C) Arrumando a variável M8-  Realização de  exame de mama
#Categorias= realizou há menos de 3 anos= 0, nunca realizou/realizou há mais de 3 anos= 1
tab1(m8)
mama<-ifelse(m8<3,1,ifelse(m8>=3,0,NA))
dados$m8 = mama
tab1(dados$m8)


## Variaveis do modelo escolhido - panorama geral:

tab1(status)  # 0 missing
tab1(tempo)   # 0 missing
tab1(idadeind) # 0 missing
tab1(corecat)  # 3.2% de missing
tab1(dados$b1)  # 0.5% de missing
tab1(dados$m7)  # 0.5% de missing
tab1(dados$m8)   # 0.5% de missing
tab1(dados$escolaridade) # 1.1% de missing
tab1(e39)  # 0.5% de missing
```

```
# Excluindo o NA das variaveis do modelo que contem missing:
banco <- dados [complete.cases (dados$b1, dados$m7, dados$m8, dados$corecat, dados$escolaridade, dados$esc1, dados$esc2,dados$esc3,dados$e39),]
dim(banco) # Tem aproximadamente 5,5% de missing em relacao a 1593.

# Removendo objetos da area de trabalho:
rm(esc1,esc2,esc3,idadeterc1,idadeterc2,idadeterc3,id)

detach(dados)
attach(banco)

## Criando banco com valor a ser retirado de cada padrao para 5, 10, 20 e 30% de missing:
# Lembrar que o % de missing eh em cima de 1505 e nao de 1593.

# Selecionando os padroes:
banco$intercepto = 1
banco2 = subset(banco, select=c("intercepto","esc2","esc3","idadeterc2","idadeterc3"))

# Colapsando os padroes para facilitar:
banco2$stringpad=apply(banco2,1,paste,collapse="")
bancopadroes= unique(banco2)
bancopadroes  # Os padroes NÃO batem certinho com a ordem do modelo anterior.

# Colocando os stringpad no banco maior:
banco$stringpadrao = banco2$stringpad

# Para resolver: unir este banco com o banco original de padroes para n=1593 (bancoprob)
# para só depois aplicar as probs:

bancopad = merge(bancoprob,bancopadroes,by="stringpad")
bancopad

# Selecionar apenas as colunas que interessam:
bancopad2=subset(bancopad,select=c("intercepto","as.factor(esc2)1","as.factor(esc3)1","as.fa
ctor(idadeterc2)1","as.factor(idadeterc3)1","stringpad","prob"))
bancopad2

# Renomeando as variáveis e criando um banco comparável para contagem do n em cada padrão:
names(bancopad2) = c("intercepto", "esc2", "esc3", "idadeterc2", "idadeterc3", "stringpad", "prob")
bancopad2

bancopad3 = subset (bancopad2, select = c("intercepto"," esc2", "esc3", "idadeterc2", "idadeterc3", "stringpad"))
bancopad3
```

```
### Verificando a quantidade de individuos presente em cada padrao:

npadrao1=integer(nrow(bancopad3))
for(j in 1:nrow(bancopad3))
        #for(i in 1:nrow(banco2))

        npadrao1[j]=nrow(subset(banco2,apply(banco2,1,paste,collapse='')==paste(bancopad3
[j,],collapse='')))

# Acrescentando o n de cada padrão ao bancopad2
bancopad4=cbind(bancopad2,npadrao1)
bancopad4

# Para saber quantos individuos devem ser removidos em cada padrao - aplicar a
probabilidade (Prob) calculada para o modelo
# de regressao logistica que rodei no banco com missing anteriormente.

m_1505=bancopad4$npadrao*bancopad4$prob

#Arredondando m para cima:
m_1505=ceiling(m_1505)
sum(m_1505)  # gera 48 indivíduos

# Juntando as informações que precisamos:
id= seq(1,9,1)
miss_padrao= cbind(id,bancopad4,m_1505)
miss_padrao   ## Banco com id, intercepto, dummies, stringpad, prob, n padrao e m_1505


# Salvando o banco com a escolaridade limpa,com as variaveis dummy do modelo para MAR
e COM a string do padrao:

bancofim=subset(banco,select=c("id","numquest","p2quest","b1","corraca","e35","e39","m7"
,"m8","idadeind","corecat",
"tercildeidade", "idadeterc1", "idadeterc2", "idadeterc3", "escolaridade", "esc1", "esc2",
"esc3", "mioma", "status", "tempo.ini", "tempo.fim", "tempo", "intercepto", "stringpadrao"))

write.csv2(bancofim,file="analise_1505_MAR.csv",row.names=F,na="")


### Criando uma lista com a quantidade de individuos a ser removido em cada padrao, para
gerar 5, 10, 20 e 30% de missing.


m_1505
sum(m_1505)  # M_1505 gera 48 individuos a serem removidos
```

```
# 48 = 3.19% de 1505

# Para gerar 5% - total =  76
# Para gerar 10% - total = 151
# Para gerar 20% - total = 301
# Para gerar 30% - total = 452

# Para ficar mais exato, ao inves de multiplicar por fatores arredondados, preferi arredondar apenas ao final, na hora de calcular a quantidade de cada padrao:

# 5%:
miss5=m_1505*(((5*1505)/100)/48)
miss5

#Arredondando:
miss5=round(miss5)
miss5
sum(miss5)  ## 76 indivíduos - 5.04% de missing

# 10% :
miss10=m_1505*(((10*1505)/100)/48)
miss10

#Arredondando:
miss10=round(miss10)
miss10
sum(miss10)  ##151 individuos - 10.038% de missing

# 20% :
miss20=m_1505*(((20*1505)/100)/48)
miss20

#Arredondando:
miss20=round(miss20)
miss20
sum(miss20)  ## 301 individuos - 20.00% de missing

# 30%:
miss30=m_1505*(((30*1505)/100)/48)
miss30

#Arredondando:
miss30=round(miss30)
miss30
sum(miss30)  ## 451 indiv?duos - 29.97% de missing
```

```
## Salvando ids dos padroes, probabilidade de cada um e os missings:
missingMAR = cbind(miss_padrao,miss5,miss10,miss20,miss30)
colnames (missingMAR) = c("id do padrao", "intercepto", "esc2", "esc3", "idadeterc2", "idadeterc3", "stringpadrao", "probabilidade", "npadrao", "m_1505", "m_5%", "m_10%", "m_20%", "m_30%")
missingMAR

write.csv2(missingMAR,file="missing_MAR.csv",row.names=F)
```

MNAR

```
# Banco a ser utilizado:  analise_1593_COX # n=1593

setwd("C:\\Documents and Settings\\Owner\\My Documents\\Dropbox\\Thais\\Análises Dissertação p. Washington\\Simulação e imputação")
dados <-read.csv2("analise_1593_23-01_COX.csv")
attach(dados)
require(epicalc)

# Avaliação dos missings da variável corecat:

# 1) Criando dummy de missing dela:

tab1(corecat)
miss = ifelse(!is.na(corecat),0,1) # 1 se for missing; 0 se não
tab1(Demissie, et al.)      # 51 missings

# Decidimos utilizar apenas as variáveis escolaridade e idade (em tercis) e, para gerar MNAR, incluir a cor/raça no modelo. COR/RAÇA - entrará com 4 categorias

# 2) Recategorizando escolaridade:
#Nome da variável nas análises= esc
#Categorias=  universitário completo= 2, 2º grau completo= 1,até 1º grau completo= 0
tab1(e35)
esc<-ifelse(e35<4,0,ifelse(e35>=4&e35<6,1,ifelse(e35>=6,2,NA)))
tab1(esc)
dados$escolaridade = esc

# 3) Recategorizando idade (tercis):
tercil = quantile(idadeind,probs=c(1/3,2/3))
tercil
```

```
idadeterc = ifelse(idadeind<=36,"tercil1",ifelse(idadeind>36&idadeind<=43,"tercil2",ifelse(idadeind>43,"tercil3",NA)))
tab1(idadeterc)
dados$tercildeidade = idadeterc

### Avaliando escolaridade (esc) e idade (tercis) de acordo com ser missing ou não:

#a) Fazendo as análises bivariadas:

## Escolaridade:
tabpct(esc,miss)   # Alguma tendência de ser mais alto em níveis menores de escolaridade

## Idade:
tabpct(idadeterc,miss) # Aparentemente,  mais missing entre os mais velhos.

## Criando dummies para idade, escolaridade e cor/raça:

# IDADE:
idadeterc1=ifelse(idadeterc=="tercil1",1,0)
dados$idadeterc1 = idadeterc1

idadeterc2=ifelse(idadeterc=="tercil2",1,0)
dados$idadeterc2 = idadeterc2

idadeterc3=ifelse(idadeterc=="tercil3",1,0)
dados$idadeterc3 = idadeterc3

# ESCOLARIDADE:

esc1 = ifelse(esc==0,1,0) # Até 1º grau completo
dados$esc1 = esc1

esc2 = ifelse(esc==1,1,0) # Até 2º grau completo
dados$esc2 = esc2
esc3 = ifelse(esc==2,1,0) # Univ completo ou mais
dados$esc3 = esc3

# COR/RAÇA:
#Observação:
# Para a variável CORECAT (pergunta aberta) = Categorias=  branca=1, parda=2, preta= 3 e amarela = 4.
# Para a variável cor/raça do IBGE (E39) - preta=1, parda=2, branca=3,amarela=4, indígena=5
# Esta variável será usada para as análises e geração dos padrões; para imputação, será CORECAT.
# Haverá um agrupamento em brancas e não brancas nesta variável (E39)
```

```
racaibge = e39
racanbranca = ifelse(racaibge==1|racaibge==2|racaibge==4|racaibge==5,1,ifelse
(racaibge==3,0,NA))
tab1(racanbranca) #conferência
dados$racanbranca = racanbranca


# MODELO MULTIVARIADO:

mod1=glm(miss~ as.factor(esc2) + as.factor(esc3) + as.factor(idadeterc2) +
as.factor(idadeterc3) +
as.factor(racanbranca), family="binomial")
summary(mod1)

# Selecionando os padrões
banco_modelo=model.matrix(mod1)
padroes=unique(banco_modelo)     ## São 18 padrões diferentes

# Transformando em data frame para o apply rodar direito:
padroes = as.data.frame(padroes)

# Colapsando os padroes para facilitar:
padroes$stringpad=apply(padroes,1,paste,collapse="")
padroes

# Identificando os padroes:
id = seq(1,18,1)
padrao=cbind(id,padroes)
padrao
### Verificando probabilidade de missing em cada padrao:

# Antes - precisei rodar novamente esta parte pq %% não aceita dataframe
banco_modelo=model.matrix(mod1)
padroes=unique(banco_modelo)

# Agora sim, a prob de missing:
preditor=padroes %*% coef(mod1)
prob=exp(preditor)/(1+exp(preditor))

### Contando quantos individuos devem ser removidos de cada padrao, no banco com
missing (n=1593)

npadrao=integer(nrow(padroes))
for(j in 1:nrow(padroes))
  #for(i in 1:nrow(banco_modelo))
```

```
npadrao[j]=nrow(subset(banco_modelo,apply(banco_modelo,1,paste,collapse='')==paste(padr
oes[j,],collapse='')))

padrao=cbind(padrao,npadrao) # acrescentar a quantidade de pessoas que há em cada padrão

m_1593=npadrao*prob

#Arredondando m para cima:
m_1593=ceiling(m_1593)

miss_padrao= cbind(id,prob,m_1593)
miss_padrao  # Banco com os "m" baseados em 1593

# Criando um banco com os padroes separados, colapsados e a prob associada a cada um:
bancoprob=cbind(padrao,prob)
bancoprob  ## Listagem de padrões, com id, string e probabilidades de cada um.


## Baseando m em 1505 (n do banco final)

# Organizando algumas variáveis do banco de dados:

# A) Arrumando a variavel plano de saude (dicotomizando) # Categorias= sim=1, nao= 0
tab1(b1)
plano<-ifelse(b1<3,1,ifelse(b1==3,0,NA))
table(plano)
dados$b1=plano  # substituir no banco para facilitar as coisas

# B) Arrumando a variável M7 - Realização de Papanicolaou
#Categorias= realizou há menos de 3 anos= 0, nunca realizou/realizou há mais de 3 anos= 1
tab1(m7)
papa<-ifelse(m7<3,1,ifelse(m7>=3,0,NA))
dados$m7 = papa
tab1(dados$m7)

## C) M8-  Realização de  exame de mama
#Categorias= realizou há menos de 3 anos= 0, nunca realizou/realizou há mais de 3 anos= 1
tab1(m8)
mama<-ifelse(m8<3,1,ifelse(m8>=3,0,NA))
dados$m8 = mama
tab1(dados$m8)

# Excluindo o NA das variáveis do modelo que contém missing (incluindo a pergunta sobre
cor/raça do IBGE (E39):
Banco =  dados [complete.cases (dados$b1, dados$m7, dados$m8, dados$corecat, dados$esc,
dados$esc1, dados$esc2, dados$esc3, dados$e39),]
```

```
dim(banco) # Tem aproximadamente 5,5% de missing em relação a 1602.

# Removendo objetos da área de trabalho:
rm(esc1,esc2,esc3,idadeterc1,idadeterc2,idadeterc3,racanbranca,id)

detach(dados)
attach(banco)

## Criando banco com valor a ser retirado de cada padrão para 5, 10, 20 e 30% de missing:
# Lembrar que o % de missing é em cima de 1505 e não de 1593.

# Selecionando os padrões:
banco$intercepto = 1
banco2 = subset (banco, select=c("intercepto", "esc2", esc3", "idadeterc2", "idadeterc3",
"racanbranca"))

# Colapsando os padroes para facilitar:
banco2$stringpad=apply(banco2,1,paste,collapse="")
bancopadroes= unique(banco2)
bancopadroes

# Colocando a string de padroes no banco maior:
banco$stringpadrao = banco2$stringpad

# Para ter certeza que a ordem de padrões é a mesma:
# unir este banco com o banco original de padroes para n=1593 (bancoprob)
# para só depois aplicar as probs:

bancopad = merge(bancoprob,bancopadroes,by="stringpad")
bancopad

# Selecionar apenas as colunas que interessam:
bancopad2=subset(bancopad,select=c("intercepto","as.factor(esc2)1","as.factor(esc3)1","as.fa
ctor(idadeterc2)1","as.factor(idadeterc3)1","as.factor(racanbranca)1","stringpad","prob"))
bancopad2

# Renomeando as variáveis e criando um banco comparável para contagem do n em cada
padrão:
names(bancopad2)                                                                =
c("intercepto","esc2","esc3","idadeterc2","idadeterc3","racanbranca","stringpad","prob")
bancopad2

bancopad3                                                                       =
subset(bancopad2,select=c("intercepto","esc2","esc3","idadeterc2","idadeterc3","racanbranca
","stringpad"))
bancopad3
```

Verificando a quantidade de indivíduos presente em cada padrão:

```
npadrao1=integer(nrow(bancopad3))
for(j in 1:nrow(bancopad3))
        #for(i in 1:nrow(banco2))
        npadrao1[j]=nrow(subset(banco2,apply(banco2,1,paste,collapse='')==paste(bancopad3
[j,],collapse='')))
# Acrescentando o n de cada padrão ao bancopad2
bancopad4=cbind(bancopad2,npadrao1)
bancopad4

# Para saber quantos indivíduos devem ser removidos em cada padrão - aplicar a
probabilidade (Prob) calculada para o modelo
# de regressão logística que rodei no banco com missing anteriormente.

m_1505=bancopad4$npadrao1*bancopad4$prob

#Arredondando m para cima:
m_1505=ceiling(m_1505)
sum(m_1505)  # Resulta em 55 individuos

# Juntando as informações que precisamos:
id= seq(1,18,1)
miss_padrao= cbind(id,bancopad4,m_1505)
miss_padrao       ## Banco com id, intercepto, dummies, stringpad, prob, n padrao e m_1505


# Salvando o banco com a escolaridade limpa e com as variáveis DUMMY do modelo para
MAR:

bancofim=subset(banco,select=c("id","numquest","p2quest","b1","corraca","e35","e39","m7"
,"m8","idadeind","corecat", "tercildeidade", "idadeterc1", "idadeterc2", "idadeterc3",
"escolaridade", "esc1","esc2","esc3","racanbranca","mioma",
"status", "tempo.ini", "tempo.fim","tempo","intercepto","stringpadrao"))

write.csv2(bancofim,file="analise_1505_MNAR.csv",row.names=F,na="")


### Criando uma lista com a quantidade de indivíduos a ser removido em cada padrão, para
gerar 5, 10, 20 e 30% de missing.

m_1505
sum(m_1505)  # M_1505 gera 55 indivíduos a serem removidos

# 55 = 3.65% de 1505
# Para ficar mais exato, ao invés de multiplicar por fatores arredondados, preferi arredondar
apenas ao final, na hora de calcular a quantidade de cada padrão:
```

```
# 5%:
miss5=m_1505*(((5*1505)/100)/55)
miss5

#Arredondando:
miss5=round(miss5)
miss5
sum(miss5)  ## 74 indivíduos - 4.92% de missing


# 10% :
miss10=m_1505*(((10*1505)/100)/55)
miss10

#Arredondando:
miss10=round(miss10)
miss10
sum(miss10)  ##147 indivíduos - 9.77% de missing


# 20% :
miss20=m_1505*(((20*1505)/100)/55)
miss20

#Arredondando:
miss20=round(miss20)
miss20
sum(miss20)  ## 299 indivíduos - 19.9% de missing

# 30%:
miss30=m_1505*(((30*1505)/100)/55)
miss30

#Arredondando:
miss30=round(miss30)
miss30
sum(miss30)  ## 450 indivíduos - 29.9% de missing

## Salvando ids dos padrões, probabilidade de cada um e os missings:
missingMNAR = cbind(miss_padrao,miss5,miss10,miss20,miss30)
colnames(missingMNAR) =c("id do padrao", "intercepto", "esc2", "esc3", "idadeterc2",
"idadeterc3", "racanbranca", "stringpadrao", "probabilidade", "npadrao", "m_1505", "m_5%",
"m_10%", "m_20%","m_30%")
write.csv2(missingMNAR,file="missing_MNAR.csv",row.names=F)
```

Modelos das funções utilizadas para replicação da simulação e análise de observações completas

MCAR

```
setwd("C:\\Documents and Settings\\Owner\\My Documents\\Dropbox\\Thais\\Análises Dissertação p. Washington\\Simulação e imputação\\MCAR")
banco=read.csv2("analise_1505_MCAR.csv")  # banco de analise

# Carregando pacotes:
require(survival)
require(epicalc)

# Selecionando apenas as variaveis de interesse no banco maior, para organizar a imputação
banco2=subset(banco,select=c("id","b1","m7","m8","idadeind","corecat","escolaridade","tercildeidade","status","tempo"))
dim(banco2)
attach(banco2)

# Simulaçao dos missings via amostra aleatoria simples dos ids, sem reposição

set.seed(1)
nrep=100
betar2=double(nrep)
betar3=double(nrep)
beta.ser2=double(nrep)
beta.ser3=double(nrep)
for(j in 1:nrep)
{

 banco_miss=banco2
 a = sample(banco_miss$id,percent) # percent = n a ser removido para gerar aquele percentual de DF
 banco_miss$corecat[banco_miss$id%in%a] <- NA

 banco_miss$corecat=as.factor(banco_miss$corecat) # Transformando cor/raca em fator

 y<-Surv(banco_miss$tempo,banco_miss$status)  # Analisar cada banco incompleto com o modelo de analise estabelecido a priori
 reg1 <- coxph (y ~ as.factor (corecat) + as.factor (tercildeidade) + as.factor (m7) + as.factor (m8)+ as.factor(b1) + as.factor (escolaridade), data=banco_miss, control = coxph.control (iter.max=100))
 sumario <- summary(reg1)
 betar2[j]<- sumario$coef[1,1]
 betar3[j] <- sumario$coef[2,1]
```

```
 beta.ser2[j] <- sumario$coef[1,3]
 beta.ser3[j] <- sumario$coef[2,3]

}# Fecho a repetição 100x

# Calculando a média e variâncias dos betas e média dos erros

# Beta 2:
# Média:
beta2 = mean(betar2)
beta2
se.beta2 = mean(beta.ser2)
se.beta2

# Variância:
varb2 = var(betar2)

# Beta 3:
# Média
beta3 = mean(betar3)
beta3
se.beta3 = mean(beta.ser3)
se.beta3

# Variância:
varb3 = var(betar3)

# Criando vetor de betas e erros:
beta = c(beta2, beta3)
se.beta =c (se.beta2, se.beta3)
var.beta = c(varb2, varb3)

# Determinando HR e ICS:
hr <- exp(beta)
liminf <- exp(beta-1.96*se.beta)
limsup <- exp(beta+1.96*se.beta)
z <- beta / se.beta
pvalor <- (1 - pnorm(abs(z), mean=0, sd=1, lower.tail = TRUE, log.p = FALSE)) * 2
data.frame( "Beta" = beta, "Erro-padrão_beta" = se.beta, "Variância" = var.beta,
"Hazard_ratio" = hr, "p-valor" = pvalor, "LI_IC95" =liminf, "LS_IC95" =limsup )
```

MAR

```
setwd("C:\\Documents and Settings\\Owner\\My Documents\\Dropbox\\Thais\\Análises
Dissertação p. Washington\\Simulação e imputação\\MAR")
banco=read.csv2("analise_1505_MAR.csv")  # banco de analise
```

```
padrao = read.csv2("missing_MAR.csv") # banco de padroes e m

# Carregando pacotes:
require(survival)
require(epicalc)

# Selecionando apenas as variaveis de interesse no banco maior, para organizar a imputação
banco2=subset(banco,select=c("id","b1","m7","m8","idadeind","corecat","escolaridade","terc
ildeidade","status","tempo","stringpadrao"))
dim(banco2)

## Simulação e análise

set.seed(1)
nrep=100
betar2=double(nrep)
betar3=double(nrep)
beta.ser2=double(nrep)
beta.ser3=double(nrep)
for(j in 1:nrep)
{
 banco_miss=NULL
 for (i in 1:nrow(padrao))
 {
  temp_i <- subset(banco2,stringpadrao==padrao[i,7])
  temp_i[sample(1:nrow(temp_i),padrao[i,'nn']),'corecat'] <- NA  ## coluna do n a ser
removido naquele padrão
  banco_miss <- rbind(banco_miss,temp_i)
 }     # Fecho este for para gerar missing

y<-Surv(banco_miss$tempo,banco_miss$status)  # Analisar cada banco incompleto com o
modelo de analise estabelecido a priori
reg1 <- coxph (y ~ as.factor (corecat) + as.factor (tercildeidade) + as.factor (m7) + as.factor
(m8) + as.factor (b1) + as.factor (escolaridade), data=banco_miss,  control = coxph.control
(iter.max=100))
a <- summary(reg1)
betar2[j]<- a$coef[1,1]
betar3[j] <- a$coef[2,1]
beta.ser2[j] <- a$coef[1,3]
beta.ser3[j] <- a$coef[2,3]

}# Fecho a repetição 100x

# Calculando a média e variâncias dos betas e média dos erros

# Beta 2:
# Média:
```

```
beta2 = mean(betar2)
beta2
se.beta2 = mean(beta.ser2)
se.beta2

# Variância:
varb2 = var(betar2)

# Beta 3:
# Média
beta3 = mean(betar3)
beta3
se.beta3 = mean(beta.ser3)
se.beta3

# Variância:
varb3 = var(betar3)

# Criando vetor de betas e erros:
beta = c(beta2, beta3)
se.beta =c (se.beta2, se.beta3)
var.beta = c(varb2, varb3)

# Determinando HR e ICS:
hr <- exp(beta)
liminf <- exp(beta-1.96*se.beta)
limsup <- exp(beta+1.96*se.beta)
z <- beta / se.beta
pvalor <- (1 - pnorm(abs(z), mean=0, sd=1, lower.tail = TRUE, log.p = FALSE)) * 2
data.frame("Beta" = beta, "Erro-padrão_beta"= se.beta,"Variância" = var.beta,"Hazard_ratio"
= hr, "p-valor" = pvalor,"LI_IC95"=liminf,"LS_IC95"=limsup)
```

MNAR

```
setwd("C:\\Documents and Settings\\Owner\\My Documents\\Dropbox\\Thais\\Análises
Dissertação p. Washington\\Simulação e imputação\\MNAR")
banco=read.csv2("analise_1505_MNAR.csv")  # banco de analise
padrao = read.csv2("missing_MNAR.csv") # banco de padroes e m

# Carregando pacotes:
require(survival)

# Selecionando apenas as variaveis de interesse no banco maior, para organizar a imputação
banco2=subset(banco,select=c("id","b1","m7","m8","idadeind","corecat","escolaridade","terc
ildeidade","status","tempo","stringpadrao"))
dim(banco2)
```

```
## Simulação

set.seed(1)
nrep=100
betar2=double(nrep)
betar3=double(nrep)
beta.ser2=double(nrep)
beta.ser3=double(nrep)
for(j in 1:nrep)
{
  banco_miss=NULL
  for (i in 1:nrow(padrao))
  {
    temp_i <- subset(banco2,stringpadrao==padrao[i,8])
    temp_i[sample(1:nrow(temp_i),padrao[i,'nn.']),'corecat'] <- NA    # nn= coluna com o n a ser removido em cada padrão
    banco_miss <- rbind(banco_miss,temp_i)
  }      # Fecho este for para gerar missing

 y<-Surv(banco_miss$tempo,banco_miss$status)  # Analisar cada banco incompleto com o modelo de analise estabelecido a priori
 reg1 <- coxph(y ~ as.factor (corecat) + as.factor (tercildeidade) + as.factor (m7) + as.factor (m8) +as.factor (b1) + as.factor (escolaridade) , data = banco_miss, control = coxph.control (iter.max=100))
a <- summary(reg1)
betar2[j]<- a$coef[1,1]
betar3[j] <- a$coef[2,1]
beta.ser2[j] <- a$coef[1,3]
beta.ser3[j] <- a$coef[2,3]

}# Fecho a repetição 100x

# Calculando a média e variâncias dos betas e média dos erros

# Beta 2:
# Média:
beta2 = mean(betar2)
beta2
se.beta2 = mean(beta.ser2)
se.beta2

# Variância:
varb2 = var(betar2)


# Beta 3:
```

```
# Média
beta3 = mean(betar3)
beta3
se.beta3 = mean(beta.ser3)
se.beta3

# Variância:
varb3 = var(betar3)

# Criando vetor de betas e erros:
beta = c(beta2, beta3)
se.beta =c (se.beta2, se.beta3)
var.beta = c(varb2, varb3)

# Determinando HR e ICS:
hr <- exp(beta)
liminf <- exp(beta-1.96*se.beta)
limsup <- exp(beta+1.96*se.beta)
z <- beta / se.beta
pvalor <- (1 - pnorm(abs(z), mean=0, sd=1, lower.tail = TRUE, log.p = FALSE)) * 2
data.frame("Beta" = beta, "Erro-padrão_beta"= se.beta,"Variância" = var.beta,"Hazard_ratio"
= hr, "p-valor" = pvalor, "LI_IC95"=liminf,"LS_IC95"=limsup)
```

Modelos das funções utilizadas para replicação da simulação, imputação e análise de dados

MCAR

```
setwd("C:\\Documents and Settings\\Owner\\My Documents\\Dropbox\\Thais\\Análises
Dissertação p. Washington\\Simulação e imputação\\MCAR")
banco=read.csv2("analise_1505_MCAR.csv")  # banco de analise

# Carregando pacotes:
require(survival)
require(mice)

# Selecionando apenas as variaveis de interesse no banco maior, para organizar a imputação
banco2=subset(banco,select=c("id","b1","m7","m8","idadeind","corecat","escolaridade","terc
ildeidade","status","tempo"))
dim(banco2)
attach(banco2)

# Simulaçao dos missings via amostra aleatoria simples dos ids, sem reposição

set.seed(1)
```

```
imput = número de bancos a ser criado com a imputação
nrep=100
betar2=double(nrep)
betar3=double(nrep)
beta.ser2=double(nrep)
beta.ser3=double(nrep)
for(j in 1:nrep)
{

 banco_miss=banco2
 a = sample(banco_miss$id,percent) # percent = n a ser removido para gerar o % de DF desejado
 banco_miss$corecat[banco_miss$id%in%a] <- NA

 banco_miss$corecat=as.factor(banco_miss$corecat) # Transformando cor/raca em fator

  ini <- mice(banco_miss, maxit=0, pri=F) # Cria as imputacoes
 pred <- ini$pred # Tirar variaveis da matriz de predicao e nao imputar as que nao interessam
 pred[,c("id","idadeind")] <- 0      # Retirei da predicao essas variaveis e deixei apenas b1 - plano de saude, m7 e m8 - exames, escolaridade recateg, tercil de idade e status e o tempo (desfechos do Cox).
 imp <- mice(banco_miss,m=imput, meth=c("","","","","","","polyreg","","","","",""), pred=pred)

 y<-Surv(banco_miss$tempo,banco_miss$status)  # Analisar cada banco completo com o modelo de analise estabelecido a priori
 fit <- with(imp, coxph(y ~ corecat + as.factor (tercildeidade) + as.factor (m7) + as.factor (m8) + as.factor (b1) + as.factor (escolaridade) , data = banco_miss, control = coxph.control (iter.max=100)))
 a <- summary(pool(fit)) # Combinar as analises usando as regras de Rubin
 betar2[j]<- a[1,1]
 betar3[j] <- a[2,1]
 beta.ser2[j]<- a[1,2]
 beta.ser3[j]<- a[2,2]

} # Fecho a repetição 100x

# Calculando a média e variâncias dos betas e média dos erros

# Beta 2:
# Média:
beta2 = mean(betar2)
beta2
se.beta2 = mean(beta.ser2)
se.beta2

# Variância:
varb2 = var(betar2)
```

```
# Beta 3:
# Média
beta3 = mean(betar3)
beta3
se.beta3 = mean(beta.ser3)
se.beta3

# Variância:
varb3 = var(betar3)

# Criando vetor de betas e erros:
beta = c(beta2, beta3)
se.beta =c (se.beta2, se.beta3)
var.beta = c(varb2, varb3)

# Determinando HR e ICS:
hr <- exp(beta)
liminf <- exp(beta-1.96*se.beta)
limsup <- exp(beta+1.96*se.beta)
z <- beta / se.beta
pvalor <- (1 - pnorm(abs(z), mean=0, sd=1, lower.tail = TRUE, log.p = FALSE)) * 2
data.frame("Beta" = beta, "Erro-padrão_beta"= se.beta,"Variância" =
var.beta,"Hazard_ratio"=hr,"p-valor"=pvalor,"LI_IC95"=liminf,"LS_IC95"=limsup)
```

MAR

```
setwd("C:\\Documents and Settings\\Owner\\My Documents\\Dropbox\\Thais\\Análises
Dissertação p. Washington\\Simulação e imputação\\MAR") # Note
banco=read.csv2("analise_1505_MAR.csv")  # banco de analise
padrao = read.csv2("missing_MAR.csv") # banco de padroes e m

# Carregando pacotes:
require(mice)
require(survival)

# Selecionando apenas as variaveis de interesse no banco maior, para organizar a imputação
banco2=subset(banco,select=c("id","b1","m7","m8","idadeind","corecat","escolaridade","terc
ildeidade","status","tempo","stringpadrao"))
dim(banco2)
## Função:

set.seed(1)
nrep=100
nn = banco2$ coluna com o n a ser removido em cada padrão
imput = número de bancos de dados a ser criado com a imputação
```

```
betar2=double(nrep)
betar3=double(nrep)
beta.ser2=double(nrep)
beta.ser3=double(nrep)
for(j in 1:nrep)
{
 banco_miss=NULL
 for (i in 1:nrow(padrao))
 {
  temp_i <- subset(banco2,stringpadrao==padrao[i,7])
  temp_i[sample(1:nrow(temp_i),padrao[i,nn.']),'corecat'] <- NA
  banco_miss <- rbind(banco_miss,temp_i)
 }     # Fecho este for para gerar missing

 banco_miss$corecat=as.factor(banco_miss$corecat) # Transformando cor/raca em fator

 ini <- mice(banco_miss, maxit=0, pri=F) # Cria as imputacoes - para m=5
 pred <- ini$pred # Tirar variaveis da matriz de predicao e nao imputar as que nao interessam
 pred[,c("id","idadeind","stringpadrao")] <- 0     # Retirei da predicao essas variaveis e deixei
apenas b1 - plano de saude, m7 e m8 - exames, escolaridade recateg, tercil de idade e status e
o tempo (desfechos do Cox).
 imp     <-     mice(banco_miss,m=imput,     meth=c("","","","","","polyreg","","","","","",""),
pred=pred)

  y<-Surv(banco_miss$tempo,banco_miss$status)  # Analisar cada banco completo com o
modelo de analise estabelecido a priori
 fit                                                                                    <-
with(imp,coxph(y~corecat+as.factor(tercildeidade)+as.factor(m7)+as.factor(m8)+as.factor(b1
)+as.factor(escolaridade),data=banco_miss, control=coxph.control(iter.max=100)))
 a <- summary(pool(fit)) # Combinar as analises usando as regras de Rubin
 betar2[j]<- a[1,1]
 betar3[j] <- a[2,1]
 beta.ser2[j]<- a[1,2]
 beta.ser3[j]<- a[2,2]

}# Fecho a repetição 100x

# Calculando a média dos betas e dos erros

# Beta 2:
beta2 = mean(betar2)
beta2
se.beta2 = mean(beta.ser2)
se.beta2

# Variância:
varbeta2 = var(betar2)
```

```
# Beta 3:
beta3 = mean(betar3)
beta3
se.beta3 = mean(beta.ser3)
se.beta3

# Variância:
varbeta3 = var(betar3)

# Criando vetor de betas, erros e variâncias:
beta = c(beta2, beta3)
se.beta =c (se.beta2, se.beta3)
variancia = c(varbeta2,varbeta3)

# Determinando HR e ICS:
hr <- exp(beta)
liminf <- exp(beta-1.96*se.beta)
limsup <- exp(beta+1.96*se.beta)
z <- beta / se.beta
pvalor <- (1 - pnorm(abs(z), mean=0, sd=1, lower.tail = TRUE, log.p = FALSE)) * 2
data.frame("Beta" = beta, "Erro-padrão_beta"= se.beta,"Variância"= variancia,
"Hazard_ratio"=hr,"p-valor"=pvalor,"LI_IC95"=liminf,"LS_IC95"=limsup)
```

MNAR

```
setwd("C:\\Documents and Settings\\Owner\\My Documents\\Dropbox\\Thais\\Análises
Dissertação p. Washington\\Simulação e imputação\\MNAR")
banco=read.csv2("analise_1505_MNAR.csv")  # banco de analise
padrao = read.csv2("missing_MNAR.csv") # banco de padroes e m

# Carregando pacotes:
require(mice)
require(survival)

# Selecionando apenas as variaveis de interesse no banco maior, para organizar a imputação
banco2=subset(banco,select=c("id","b1","m7","m8","idadeind","corecat","escolaridade","terc
ildeidade","status","tempo","stringpadrao"))
dim(banco2)

# Função:

set.seed(1)
nrep=100
nn = banco2$ coluna com o n a ser removido de cada padrão
imput = número de bancos de dados a ser criado com a imputação
```

```
betar2=double(nrep)
betar3=double(nrep)
beta.ser2=double(nrep)
beta.ser3=double(nrep)
for(j in 1:nrep)
{
 banco_miss=NULL
 for (i in 1:nrow(padrao))
 {
  temp_i <- subset(banco2,stringpadrao==padrao[i,8])
  temp_i[sample(1:nrow(temp_i),padrao[i,nn]),'corecat'] <- NA
  banco_miss <- rbind(banco_miss,temp_i)
 }     # Fecho este for para gerar missing

 banco_miss$corecat=as.factor(banco_miss$corecat) # Transformando cor/raca em fator

 ini <- mice(banco_miss, maxit=0, pri=F) # Cria as imputacoes - para m=5
 pred <- ini$pred # Tirar variaveis da matriz de predicao e nao imputar as que nao interessam
 pred[,c("id","idadeind","stringpadrao")] <- 0     # Retirei da predicao essas variaveis e deixei
apenas b1 - plano de saude, m7 e m8 - exames, escolaridade recateg, tercil de idade e status e
o tempo (desfechos do Cox).
 imp <- mice(banco_miss,m=5, meth=c("","","","","","polyreg","","","","","",""), pred=pred)

  y<-Surv(banco_miss$tempo,banco_miss$status)  # Analisar cada banco completo com o
modelo de analise estabelecido a priori
 fit <- with (imp, coxph (y ~ corecat + as.factor (tercildeidade) + as.factor (m7) + as.factor
(m8) + as.factor (b1) + as.factor (escolaridade) , data = banco_miss, control = coxph.control
(iter.max=100)))
 a <- summary(pool(fit)) # Combinar as analises usando as regras de Rubin
 betar2[j]<- a[1,1]
 betar3[j] <- a[2,1]
 beta.ser2[j]<- a[1,2]
 beta.ser3[j]<- a[2,2]

}# Fecho a repetição 100x

# Calculando a média dos betas e dos erros e as variâncias

# Beta 2:
beta2 = mean(betar2)
beta2
se.beta2 = mean(beta.ser2)
se.beta2

# Variância:
varbeta2 = var(betar2)
```

```
# Beta 3:
beta3 = mean(betar3)
beta3
se.beta3 = mean(beta.ser3)
se.beta3

# Variância:
varbeta3 = var(betar3)

# Criando vetor de betas e erros:
beta = c(beta2, beta3)
se.beta =c (se.beta2, se.beta3)
variancia = c(varbeta2, varbeta3)

# Determinando HR e ICS:
hr <- exp(beta)
liminf <- exp(beta-1.96*se.beta)
limsup <- exp(beta+1.96*se.beta)
z <- beta / se.beta
pvalor <- (1 - pnorm(abs(z), mean=0, sd=1, lower.tail = TRUE, log.p = FALSE)) * 2
data.frame("Beta" = beta, "Erro-padrão_beta"= se.beta,"Variância" =
variancia,"Hazard_ratio"=hr,"p-valor"=pvalor,"LI_IC95"=liminf,"LS_IC95"=limsup)
```

Script utilizado para análise de *performance* dos diferentes cenários

```
# Carregar a área de trabalho necessária:

setwd("C:\\Documents and Settings\\Owner\\My Documents\\Dropbox\\Thais\\Análises
Dissertação p. Washington\\Simulação e imputação\\MNAR\\imputacao_OK\\5%")
load("area.Rdata")
ls()

# Valores de beta do banco original
betapad2= 0.12800
betapad3=0.58362

# Betas das simulações
betar2
betar3
# Fórmulas que precisarei para cada beta:

# Diferença entre padrão e cada beta
dif2 = betapad2 - betar2
dif2
```

```
dif3 = betapad3 - betar3
dif3

# Soma das diferenças (para viés)
somadif2 = sum(dif2)
somadif3 = sum(dif3)

# Diferenças ao quadrado (para DMQ)
dif22 = dif2^2
dif22

dif32 = dif3^2
dif32

# Soma do quadrado das diferenças:
soma2 = sum(dif22)
soma2

soma3 = sum(dif32)
soma3

# Raiz da soma:
raiz2 = sqrt(soma2)
raiz2

raiz3 = sqrt(soma3)
raiz3

# Diferença entre os 100 betas e a média dos 100 betas:
difmedia2 = betar2 - beta2
difmedia2

difmedia3 = betar3 - beta3
difmedia3

# Diferença entre o beta original e a média dos 100 betas
diforiginal2 = betapad2-beta2
diforiginal2

diforiginal3 = betapad3-beta3
diforiginal3

m=100

# MEDIDAS:

# DMQ:
```

```
dmq2 = (1/m) * (raiz2)
dmq2

dmq3 = (1/m) * (raiz3)
dmq3

# VIÉS:
vies2 = (1/m) * (somadif2)
vies2

vies3 = (1/m) * (somadif3)
vies3
```