CSP CADERNOS DE SAÚDE PÚBLICA
REPORTS IN PUBLIC HEALTH

COMUNICAÇÃO BREVE
BRIEF COMMUNICATION

Simulação e comparação de técnicas de correção de dados incompletos de idade para o cálculo de taxas de incidência

Simulation and comparison of techniques for the correction of incomplete data on age to calculate incidence rates

Simulación y comparación de técnicas de corrección de datos incompletos de edad para el cálculo de tasas de incidencia

Max Moura de Oliveira [1,2]
Maria do Rosário Dias de Oliveira Latorre [1]
Luana Fiengo Tanaka [1,3]
Maria Paula Curado [2,4]

doi: 10.1590/0102-311X00140717

Resumo

O objetivo foi comparar duas técnicas para estimar idade em bancos de dados com registros incompletos e analisar sua aplicação no cálculo da incidência de câncer. Utilizou-se a base de dados do Registro de Câncer de Base Populacional do Município de São Paulo, Brasil, contendo casos diagnosticados por câncer do trato urinário, entre 1997 e 2013. Foram aplicadas duas técnicas para estimativa de idade: fator de correção e imputação múltipla. Foram simuladas, usando a distribuição binomial, seis bases de dados com diferentes proporções de dados incompletos para idade de 5% até 50%. A razão entre as incidências foi calculada tendo, como referência, a base completa, cuja incidência padronizada foi de 11,83/100 mil; as demais incidências nas bases com 5% ou mais de dados incompletos para idade apresentaram-se subestimadas. Ao aplicar o fator de correção, as taxas corrigidas não apresentaram diferenças em comparação com as padronizadas, entretanto, essa técnica não permite corrigir taxas específicas. A imputação múltipla foi útil na correção das taxas padronizadas e específicas em bancos com até 30% de dados incompletos, entretanto, as taxas específicas para indivíduos com menos de 50 anos apresentaram-se subestimadas. Bases com 5% ou mais de dados incompletos necessitam de aplicação de correção. A imputação múltipla, apesar de complexa em sua execução, mostrou-se superior ao fator de correção. Todavia, deve ser utilizada com parcimônia, pois taxas específicas por idade podem manter-se subestimadas.

Incidência; Indicadores Básicos de Saúde; Base de Dados; Neoplasias

Correspondência
M. M. Oliveira
Registro de Câncer de Base Populacional de São Paulo, Faculdade de Saúde Pública, Universidade de São Paulo.
Av. Dr. Arnaldo 715, 1º andar, São Paulo, SP 01246-904, Brasil.
max.moura@usp.br

[1] Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo, Brasil.
[2] Centro Internacional de Pesquisa, A.C. Camargo Cancer Center, São Paulo, Brasil.
[3] Faculty of Sport and Health Sciences, Technical University of Munich, Munich, Germany.
[4] International Prevention Research Institute, Écully, France.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições, desde que o trabalho original seja corretamente citado.

Introdução

Os sistemas de informação em saúde (SIS) fornecem dados para gestão, planejamento e estudos epidemiológicos, e os indicadores mais utilizados são as taxas padronizadas e específicas por idade [1,2,3]. Apesar da melhoria da qualidade dos dados [3], os SIS ainda apresentam dados incompletos (*missing data*) [4,5,6]. Técnicas como o *linkage* trazem ganho na completude das bases de dados utilizando outras fontes, porém são dependentes da qualidade do SIS utilizado [7,8].

Em bases com dados incompletos, recomenda-se a aplicação de correção para subestimativas. O fator de correção é uma técnica, de simples aplicação, que consiste na correção da taxa padronizada pelo percentual de dados ausentes [9].

Outra técnica de correção é a imputação de dados. Inicialmente, a imputação era única e substituía os dados faltantes pela média, mediana, interpolação ou regressão linear. Posteriormente, surgiram as técnicas de imputação múltipla para corrigir incertezas associadas à imputação única, pois consideram características do próprio indivíduo ao qual terá o dado atribuído [4,10].

Dentre os SIS, os registros de câncer de base populacional (RCBP) fornecem dados sobre a incidência do câncer [9,11], com possibilidade de comparação entre localidades ou anos diferentes. Para tanto, a taxa é padronizada pela idade. Assim, se essa variável apresentar dados incompletos, as taxas estarão subestimadas [12]. Neste estudo, bases com diferentes percentuais de dados incompletos de idade foram simuladas para comparar duas técnicas de correção da incidência.

Método

Foram utilizados dados do RCBP do Município de São Paulo, Brasil, contendo 22.859 casos incidentes de câncer do trato urinário (C64-C68), no período de 1997 a 2013. A base que não apresentava dados faltantes para a idade foi denominada base zero (0). A partir da base 0, foram simuladas seis bases com diferentes percentuais de perdas para a variável idade: base 1 (5%), 2 (10%), 3 (20%), 4 (30%), 5 (40%) e 6 (50%). Os dados faltantes foram gerados assumindo distribuição binomial, pelo comando *rbimom(n,size,prob)*, do software R (The R Foundation for Statistical Computing; http://www.r-project.org).

Foram calculadas as taxas de incidência específicas por idade (< 50 anos e ≥ 50 anos) e taxas de incidência padronizada (TIP). A padronização das taxas foi pelo método direto, utilizando, como população padrão, em que são estimados os casos esperados numa determinada população, nesse caso, a população mundial [9,11,12].

Para a correção, foram comparadas duas estratégias. A primeira foi a aplicação de um fator de correção (FC) proposto por Bray e Ferlay e utilizado na publicação *Cancer Incidence in Five Continents* [9]. O FC é obtido pelo resultado da razão do número total de casos dividido pelo número de casos com idade conhecida para mesma topografia e sexo. Ou seja:

Taxa Padronizada Corrigida = Taxa de Incidência Padronizada * Fator de Correção

A segunda técnica utilizada foi a imputação múltipla. Essa foi proposta por Rubin, nos anos 1970, com intuito de solucionar o problema da não-resposta em pesquisas [4,10].

Para cada uma das seis bases, foram, inicialmente, obtidos 100 bancos (m) com dados completos por meio da aplicação da imputação múltipla, assumiram-se as perdas ao acaso (MAR – *missing at random*), e foi utilizado o método *predictive mean matching* (PMM) para estimar as idades faltantes. As variáveis incluídas no modelo preditivo para imputação foram as disponíveis no banco do RCBP, com excelente completude e relacionadas aos dados faltantes: sexo e topografia (relacionadas ao indivíduo); ano e fonte notificadora (momento do diagnóstico).

Para combinar os m bancos obtidos, foi aplicada a regra de Rubin, que é a média das estimativas realizadas para cada indivíduo e pode ser utilizada independentemente do método de imputação [5,10]. Para a realização das imputações, foram seguidas as seguintes instruções do R:

```
library(mice)
data <- read_excel("basededados")
summary(data)
md.pattern(data)
meth[c("variáveis a imputar ")]="pmm"
imputedidade = mice(data, method=meth, predictorMatrix=predM, m=100, mech="MAR")
bancoimput <- complete(imputedidade)
```

Foram calculados as TIP e as razões das taxas de incidência (*standardized incidence ratios* – SIR) e os intervalos de 95% de confiança pelo comando *rateratio.wald(c(casos referência, casos, população referência, população)*. A base 0 foi utilizada como referência. As análises das taxas e das SIR foram calculadas no pacote *epitools* 3.4.3. E a imputação múltipla, no pacote MICE (*multivariate imputation by chained equations*), ambos no software R.

Resultados

Na Figura 1, são apresentados os percentuais de perda e as idades aleatorizadas segundo cada base, evidenciando uma distribuição aleatória dessas perdas. A taxa média de incidência padronizada oriunda da base completa (0) foi de 11,83 por 100 mil habitantes. As taxas calculadas com dados provenientes de bases com 5% ou mais de dados faltantes mostraram-se estatisticamente subestimadas em todos os anos estudados. As razões das taxas foram significativamente menores que 1. Houve aumento da subestimativa de acordo com o percentual de dados incompletos (Tabela 1; Figura 2a).

O FC aplicado nas bases com os diferentes percentuais de dados faltantes foi útil para corrigir as subestimativas. As razões das taxas foram iguais a 1, ou seja, não apresentaram diferenças significativas (Tabela 1; Figura 2b).

Figura 1

Histograma dos dados faltantes e dados aleatorizados, segundo bancos de dados. Município de São Paulo, Brasil, 1997-2013.

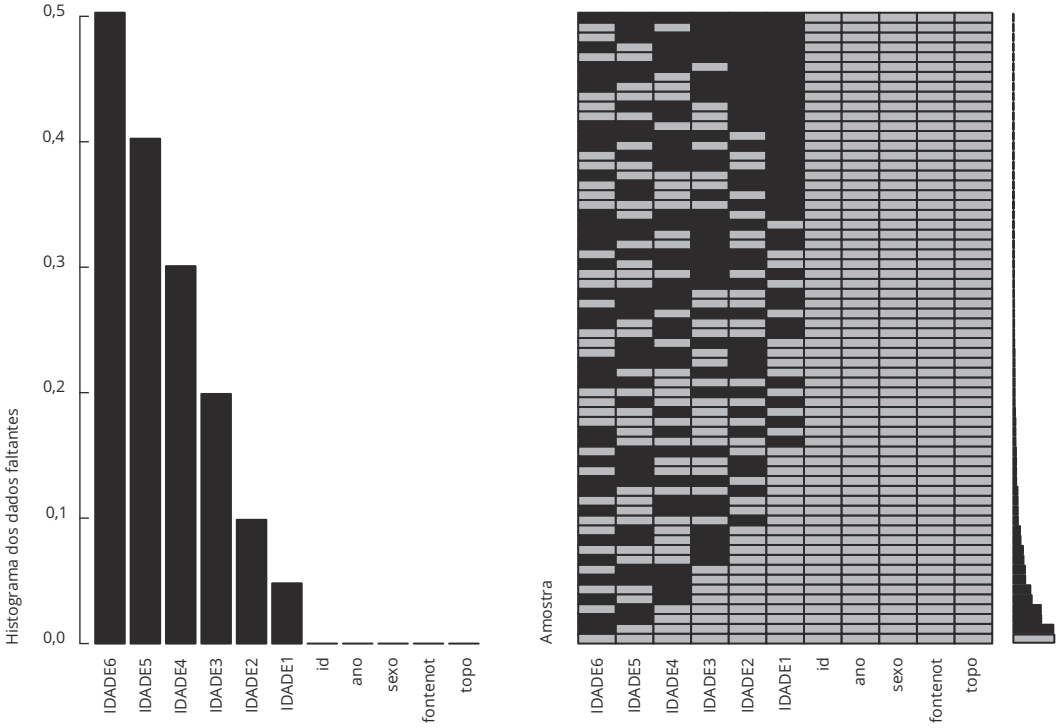


Tabela 1

Taxa de incidência padronizada e específica por idade, razões das taxas (RT), segundo bancos de dados. Município de São Paulo, Brasil, 1997-2013.

Faixa etária (anos)	Base (% de idade incompleta)	Idade incompleta			Fator de correção			Dados imputados		
		Taxa	RT	IC95%	Taxa	RT	IC95%	Taxa	RT	IC95%
Todas as idades	Base completa 0 (0%)	11,83				Referência				
	1 (5%)	11,26	0,95	0,93; 0,97	11,83	1,00	0,98; 1,01	11,77	0,99	0,97; 1,01
	2 (10%)	10,67	0,90	0,88; 0,91	11,84	1,00	0,98; 1,01	11,74	0,99	0,97; 1,01
	3 (20%)	9,47	0,80	0,78; 0,81	11,82	1,00	0,98; 1,01	11,88	1,00	0,98; 1,02
	4 (30%)	8,27	0,70	0,68; 0,71	11,83	1,00	0,98; 1,01	11,55	0,98	0,95; 0,99
	5 (40%)	7,07	0,60	0,58; 0,61	11,83	1,00	0,98; 1,01	11,41	0,96	0,94; 0,98
	6 (50%)	5,89	0,50	0,48; 0,50	11,85	1,00	0,98; 1,02	11,94	1,01	0,99; 1,02
< 50	Base completa 0 (0%)	2,25				Referência				
	1 (5%)	2,14	0,96	0,91; 1,00	-	-	-	2,16	0,96	0,92; 1,01
	2 (10%)	2,03	0,91	0,86; 0,95	-	-	-	2,09	0,93	0,89; 0,99
	3 (20%)	1,83	0,82	0,78; 0,86	-	-	-	1,88	0,84	0,80; 0,88
	4 (30%)	1,57	0,70	0,67; 0,74	-	-	-	1,73	0,77	0,74; 0,81
	5 (40%)	1,37	0,61	0,58; 0,65	-	-	-	1,53	0,68	0,65; 0,72
	6 (50%)	1,12	0,50	0,47; 0,53	-	-	-	1,36	0,61	0,58; 0,64
≥ 50	Base completa 0 (0%)	50,13				Referência				
	1 (5%)	47,72	0,95	0,93; 0,97	-	-	-	50,24	1,00	0,98; 1,02
	2 (10%)	45,23	0,90	0,88; 0,92	-	-	-	50,31	1,00	0,98; 1,02
	3 (20%)	40,00	0,80	0,78; 0,82	-	-	-	51,9	1,04	1,01; 1,06
	4 (30%)	35,04	0,70	0,68; 0,71	-	-	-	50,82	1,01	0,99; 1,03
	5 (40%)	29,90	0,60	0,58; 0,61	-	-	-	50,92	1,02	1,00; 1,04
	6 (50%)	24,97	0,50	0,49; 0,51	-	-	-	54,26	1,08	1,06; 1,10

IC95%: intervalo de 95% de confiança.

As TIP calculadas utilizando as bases com 5% a 20% de dados faltantes corrigidas com imputação múltipla não apresentaram diferenças significativas (razões das taxas igual a 1). Quando houve diferenças, essas foram menores que 5% de superestimação (Tabela 1; Figura 2c).

Ao calcular as taxas de incidência específicas por idade com bases incompletas, essas se mantiveram subestimadas. A técnica de imputação múltipla permitiu corrigir a taxa específica por idade, porém, gerou subestimativa na faixa etária mais jovem (até 50 anos) (Tabela 1).

Discussão

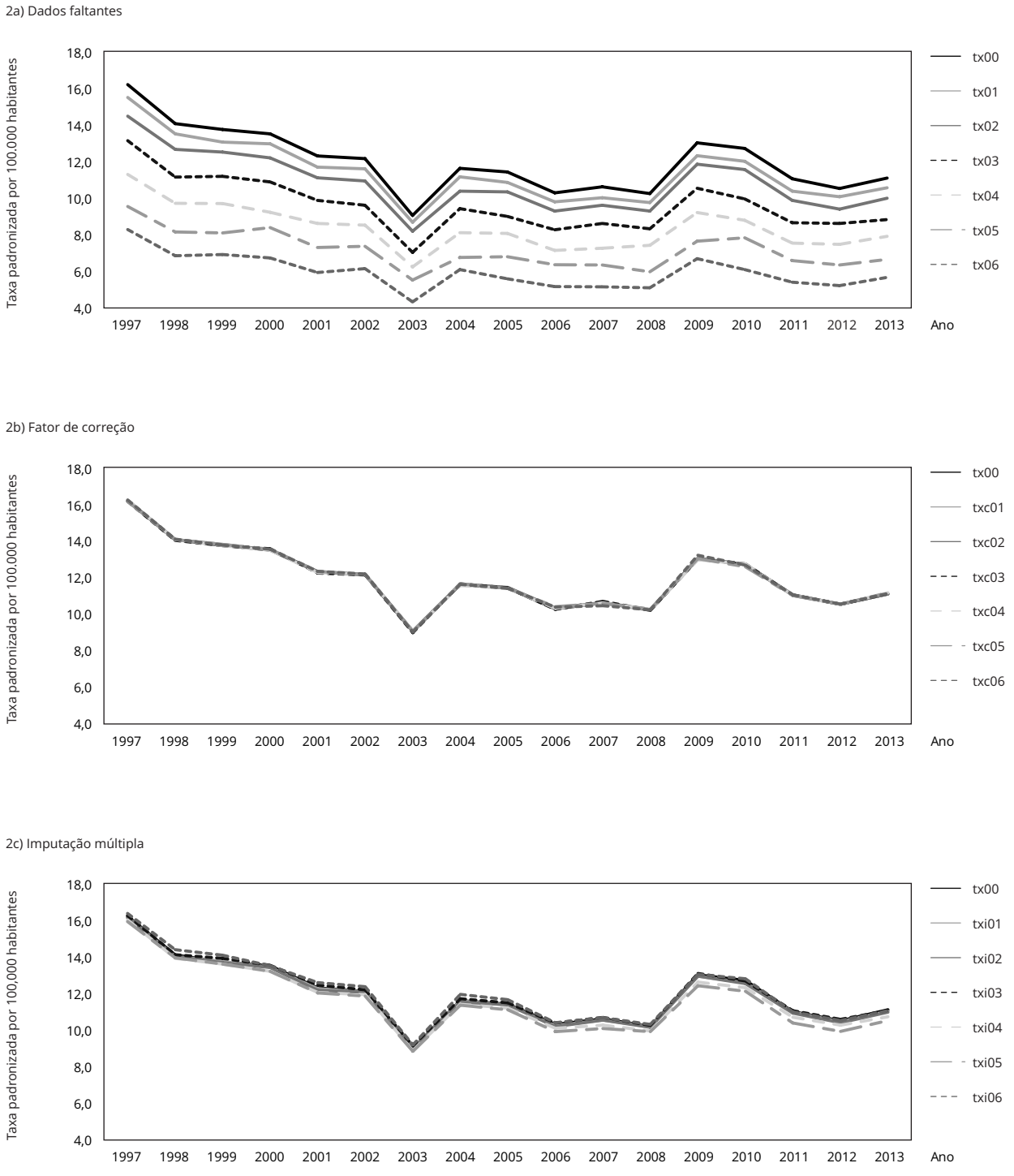
A TIP permite comparar a incidência entre diferentes populações. Estudar a subestimativa das taxas devido à idade faltante é relevante, pois é a característica mais utilizada nas padronizações de taxas, devido a diferenças nas estruturas etárias de áreas geográficas distintas [9,11,12].

A Agência Internacional de Pesquisa sobre Câncer (IARC) aponta a necessidade de corrigir taxas com ≥ 10% de dados faltantes [9]. Apesar de bastante comum, a restrição das análises com os dados completos [4,5] não é a mais adequada. Neste estudo, bases com mais de 5% de dados faltantes resultaram em taxas subestimadas.

As duas técnicas utilizadas foram úteis para mitigar a subestimativa das taxas. O FC proposto e utilizado pela IARC é uma técnica de simples execução. Essa técnica assume que a idade faltante é distribuída aleatoriamente. Ou seja, a probabilidade de a idade de um caso ser desconhecida

Figura 2

Taxa padronizada de câncer do trato urinário, segundo bancos de dados. Município de São Paulo, Brasil, 1997-2013.



tx00: taxa padronizada do banco completo (referência); tx01 a tx06: taxa padronizada dos bancos com dados faltantes; txc01 a txc06: taxa padronizada corrigida pelo fator de correção; txi01 a txi06: taxa padronizada com bancos imputados. Dados faltantes dos bancos: 1 = 5%; 2 = 10%; 3 = 20%; 4 = 30%; 5 = 40% e 6 = 50%.

independe da idade do caso. Entretanto, a probabilidade de um caso não ter a idade registrada é maior entre indivíduos mais velhos, pelo aumento do risco de câncer com a idade, por isso, todos os casos registrados devem ser considerados no cálculo para evitar subestimativas [9].

Apesar de ambas as técnicas não apresentarem diferenças, o FC é útil apenas quando o interesse é corrigir a TIP para todos os indivíduos. A imputação múltipla de dados foi mais adequada ao possibilitar o cálculo das taxas de incidência padronizadas e específicas. Todavia, taxas para indivíduos até 50 anos mantiveram-se subestimadas após imputação, o que pode estar relacionado com o maior risco do câncer entre indivíduos mais velhos [9]. O uso do PMM ocorre quando há interesse em imputar variáveis de naturezas distintas [10,13], como dados dos RCBP. A opção por m = 100 foi pela melhoria na precisão dos valores combinados, resultando em melhores estimativas [14].

A correção da TIP utilizando a técnica de imputação múltipla de dados levou, em consideração, as características tanto dos indivíduos quanto do momento da notificação, conforme a proposição do método [4,10]. O ano de diagnóstico foi utilizado devido à melhoria dos dados ao longo dos anos.

Os intervalos de confiança das razões das taxas comparando as taxas do banco completo e os obtidos pela imputação múltipla evidenciaram superestimava das taxas padronizadas em bases com mais de 30% de dados faltantes. A superestimava foi menor do que a subestimativa das taxas com o mesmo percentual de dados faltantes.

Foi possível identificar diferenças nas estratégias utilizadas, sugerindo que, em neoplasias que a incidência aumente com a idade, a imputação múltipla corrige subestimativas. Entretanto, é necessário testar a imputação múltipla em outras neoplasias, pois, para o trato urinário, pacientes de idade intermediária apresentaram incidência subestimada. Como limitação, os RCBP possuem baixa completude de outras variáveis que poderiam contribuir na precisão dos dados estimados.

Bases de dados com mais de 5% de idade ignorada apresentam taxas de incidência subestimadas. Entretanto, em bases com mais de 30% de dados faltantes, as técnicas de correção devem ser utilizadas com cautela. A imputação múltipla permitiu a correção das TIPs e das taxas de incidência específicas.

Colaboradores

M. M. Oliveira e M. R. D. O. Latorre contribuíram com a concepção e projeto, análise e interpretação dos dados, redação do artigo e revisão crítica relevante do conteúdo intelectual. L. F. Tanaka e M. P. Curado contribuíram com a redação do artigo e revisão crítica relevante do conteúdo intelectual. Todos os autores aprovaram a versão a ser publicada e são responsáveis por todos os aspectos do trabalho na garantia da exatidão e integridade de qualquer parte da obra.

Agradecimentos

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) pela concessão da bolsa de doutorado.

Referências

1. Cintho LMM, Machado RR, Moro CMC. Métodos para avaliação de sistema de informação em saúde. J Health Inform 2016; 8:41-8.
2. Nykänen P, Brender J, Talmon J, de Keizer N, Rigby M, Beuscart-Zephir M-C, et al. Guideline for good evaluation practice in health informatics (GEP-HI). Int J Med Inform 2011; 80:815-27.
3. Mello-Jorge MHP, Laurenti R, Gotlieb SLD. Avaliação dos sistemas de informação em saúde no Brasil. Cad Saúde Colet (Rio J.) 2010; 18:7-18.
4. Rubin DB. Multiple imputation after 18+ years. J Am Stat Assoc 1996; 91:473-89.
5. Nunes LN, Klück MM, Fachel JMG. Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. Cad Saúde Pública 2009; 25:268-78.
6. 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.
7. Maia LTS, Souza WV, Mendes ACG. A contribuição do linkage entre o SIM e SINASC para a melhoria das informações da mortalidade infantil em cinco cidades brasileiras. Rev Bras Saúde Matern Infant 2015; 15:57-66.
8. Peres SV, Latorre MRDO, Tanaka LF, Michels FAS, Teixeira MLP, Coeli CM, et al. Melhora na qualidade e completitude da base de dados do Registro de Câncer de Base Populacional do Município de São Paulo: uso das técnicas de linkage. Rev Bras Epidemiol 2016; 19:753-65.
9. Forman D, Bray F, Brewster D, Gombe Mbalawa C, Kohler B, Piñeros M, et al. Cancer incidence in five continents. v. X. Lyon: International Agency for Research on Cancer; 2014. (IARC Scientific Publication, 164).
10. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. J Stat Softw 2011; 45:1-67.
11. Doll R, Payne P, Waterhouse P. Cancer incidence in five continents. Berlin: Springer; 1970.
12. Gordis L. Epidemiology. Philadelphia: Elsevier Saunders; 2009.
13. Eisemann N, Waldmann A, Katalinic A. Imputation of missing values of tumour stage in population-based cancer registration. BMC Med Res Methodol 2011; 11:1-13.
14. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med 2011; 30:377-99.

Abstract

The objective was to compare two techniques to estimate age in databases with incomplete records and analyze their application to the calculation of cancer incidence. The study used the database of the Population-Based Cancer Registry from the city of São Paulo, Brazil, containing cases of urinary tract cancer diagnosed from 1997 to 2013. Two techniques were applied to estimate age: correction factor and multiple imputation. Using binomial distribution, six databases were simulated with different proportions of incomplete data on patient's age (from 5% to 50%). The ratio between the incidence rates was calculated, using the complete database as reference, whose standardized incidence was 11.83/100,000; the other incidence rates in the databases, with at least 5% incomplete data for age, were underestimated. By applying the correction factors, the corrected rates did not differ from the standardized rates, but this technique does not allow correcting specific rates. Multiple imputation was useful for correcting the standardized and specific rates in databases with up to 30% of incomplete data, but the specific rates for individuals under 50 years of age were underestimated. Databases with 5% incomplete data or more require correction. Although the implementation of multiple imputation is complex, it proved to be superior to the correction factor. However, it should be used sparingly, since age-specific rates may remain underestimated.

Incidence; Health Status Indicators; Database; Neoplasms

Resumen

El objetivo fue comparar dos técnicas para estimar edad en bancos de datos con registros incompletos y analizar su aplicación en el cálculo de la incidencia de cáncer. Se utilizó la base de datos del Registro de Cáncer de Base Poblacional del municipio de São Paulo, Brasil, conteniendo casos diagnosticados de cáncer del tracto urinario, entre 1997 y 2013. Se aplicaron dos técnicas para la estimativa de edad: factor de corrección e imputación múltiple. Fueron simuladas, usando una distribución binomial, seis bases de datos con diferentes proporciones de datos incompletos para edad desde un 5% hasta el 50%. La razón entre las incidencias se calculó teniendo, como referencia, la base completa, cuya incidencia padronizada fue de 11,83/100.000; las demás incidencias en las bases con un 5% o más de datos incompletos en la edad se presentaron subestimadas. Al aplicar el factor de corrección, las tasas corregidas no presentaron diferencias, en comparación con las estandarizadas, sin embargo, esta técnica no permite corregir tasas específicas. La imputación múltiple fue útil en la corrección de las tasas estandarizadas y específicas en bancos con hasta un 30% de datos incompletos, no obstante, las tasas específicas para individuos con menos de 50 años se presentaron subestimadas. Bases con un 5% o más de datos incompletos necesitan una aplicación de corrección. La imputación múltiple, a pesar de ser compleja en su ejecución, se mostró superior al factor de corrección. Sin embargo, debe ser utilizada con prudencia, puesto que las tasas específicas por edad pueden seguir manteniéndose subestimadas.

Incidencia; Indicadores de Salud; Base de Datos; Neoplasias

Recebido em 16/Ago/2017
Versão final reapresentada em 13/Mar/2018
Aprovado em 16/Abr/2018