Datos incompletos: una mirada crítica para su manejo en estudios sanitarios

Maylée Cañizares / Isabel Barroso / Karen Alfonso
Instituto Nacional de Higiene, Epidemiología y Microbiología (INHEM). La Habana. Cuba.

Correspondencia: Mayilée Cañizares Pérez. Instituto Nacional de Higiene, Epidemiología y Microbiología (INHEM). Infanta 1158 entre Clavel y Llinás. Centro Habana. Ciudad de La Habana. Cuba.
Correo electrónico: mcperez@yahoo.com

Recibido: 12 de mayo de 2003.
Aceptado: 20 de octubre de 2003.

(Methods for handling incomplete data in health research: a critical look)

Resumen

Objetivo: Ilustrar los procedimientos para el manejo de datos incompletos en las investigaciones sanitarias.

Métodos: Se discuten dos estrategias para el manejo de los datos incompletos: *a)* análisis de casos completos, y *b)* imputaciones, que incluye asignar la media al valor ausente, predecir el valor ausente mediante modelos de regresión e imputaciones múltiples. Para ilustrar estos procedimientos, se elabora un ejemplo en el contexto de la regresión logística con datos procedentes de la «Segunda encuesta nacional de factores de riesgo y afecciones crónicas no trasmisibles», realizada en Cuba en el año 2001.

Resultados: Al imputar por las medias y por modelos de regresión, los resultados fueron similares y se obtuvo una *odds ratio* (OR) sobrestimada por encima del 10%. El análisis de casos completos obtuvo los resultados más alejados de las estimaciones de las OR de referencia, con una diferencia del 2 hasta el 65% de variación. Los 3 métodos invirtieron la relación entre la edad y la hipertensión. Las imputaciones múltiples fueron el método que proveyó las estimaciones más cercanas a las de referencia, con una variación menor al 16%. Éste fue el único procedimiento que preservó la relación entre la edad y la hipertensión.

Conclusiones: La elección de los procedimientos para el manejo de datos incompletos resulta una tarea compleja, pues en determinadas situaciones un mismo procedimiento puede producir estimaciones precisas y en otras no. El análisis de los datos completos debe realizarse con cautela por la pérdida sustancial de información que se genera. Las imputaciones por medias y modelos de regresión producen estimaciones poco fiables bajo mecanismos MAR *(missing at random)*.

Palabras clave: Valores ausentes. Datos incompletos. Imputaciones. Casos completos. Imputaciones múltiples.

Abstract

Objective: To illustrate methods for handling incomplete data in health research.

Methods: Two strategies for handling missing data are presented: complete-case analysis and imputations. The imputations used were mean imputations, regression imputations, and multiple imputations. These strategies are illustrated in the context of logistic regression through an example using data from the «Second Cuban national survey on risk factors and non communicable disease», carried out in 2001.

Results: The results obtained via mean and regression imputation were similar. The odds ratios were overestimated by 10%. The results of complete-case analysis showed the greatest difference from the reference odds ratios, with a variation of between 2 and 65%. The three methods distorted the relationship between age and hypertension. Multiple imputations produced estimates closest to those of the reference estimates with a variation of less than 16%. This was the only procedure preserving the relationship between age and hypertension.

Conclusions: Selecting methods for handling missing data is difficult, since the same procedure can give precise estimations in certain circumstances and not in others. Complete-case analysis should be used with caution due to the substantial loss of information it produces. Mean and regression imputations produce unreliable estimates under missing at random (MAR) mechanisms.

Key words: Missing data. Incomplete data. Imputations. Case-complete analysis. Multiple imputations.

Introducción

Los investigadores en el campo de la salud pública con frecuencia se encuentran con datos incompletos *(missing data)* al llevar a cabo sus estudios. Éstos pueden aparecer en unidades completas o en ítems de algunos sujetos. Los primeros surgen cuando las personas incluidas en el diseño del estudio no desean participar o no se puede contactar con ellas mediante el mecanismo de selección establecido, mientras que el segundo aparece cuando se tiene todas las respuestas a las preguntas del cuestionario en algunos sujetos, pero para ciertas preguntas no se tiene información alguna en el resto de los individuos.

La ausencia de datos, ya sea en unidades completas o en ítems de algunos sujetos, crea sesgos potenciales en las estimaciones de los parámetros de interés. Para ciertas características importantes, los sujetos que responden pueden ser significativamente diferentes de los que no lo hacen. En estos casos, la muestra no puede reflejar adecuadamente a los sujetos que no responden. La proyección de las características individuales de los que no responden estará sustentada en las respuestas de los que participan. Si existen grandes diferencias y éstas no se consideran en el análisis, cualquier estimación o inferencia que se haga en la población no será válida[1,2].

La no respuesta en unidades completas se compensa usando pesos muestrales, que consideran la disminución del tamaño de la muestra; mientras que para los datos incompletos en ítems hay diversas maneras de tratarlos. En los años setenta, la regla general era olvidarlos, por lo que su tratamiento consistía en la eliminación de los sujetos con información incompleta. En los años ochenta se generalizó el tratamiento de los datos incompletos a través de la búsqueda de un valor que posteriormente sería asignado al dato faltante. En la década de los noventa se produjo un cambio en la filosofía del tratamiento de los datos incompletos: ya no importa buscar un valor, sino modelar la incertidumbre alrededor de él, y se comienzan a realizar las primeras imputaciones múltiples[2].

La estrategia más apropiada para el tratamiento de datos incompletos no sólo depende de los mecanismos que lo generan, sino de las tasas de no respuesta. Cuando se considera la magnitud de la no respuesta para ciertas variables, uno de las interrogantes que pueden surgir es cómo manejar la información incompleta y cuál será el efecto de estos ajustes en los procedimientos que se emplearán. En este trabajo se presentan métodos que se pueden emplear para manejar los datos incompletos al analizar estudios sanitarios. Se ilustran con un ejemplo hipotético con fines didácticos a partir de datos procedentes de la «Segunda encuesta nacional de factores de riesgo y afecciones crónicas no trasmisibles»[3], realizada en Cuba en el año 2001. Se discuten las ventajas y limitaciones del uso de cada procedimiento.

Mecanismos de la no respuesta

Para solucionar el problema del análisis estadístico con datos incompletos, es necesario identificar, en primer lugar, el mecanismo que describe la distribución de los valores ausentes y su implicación en la inferencia estadística. La clasificación de estos mecanismos, según Little y Rubin[4], se basa en la aleatoriedad con que se distribuyen los valores ausentes en las unidades. Estos autores definen 3 tipos de mecanismos: proceso completamente aleatorio (*missing completely at random,* MCAR), proceso aleatorio (*missing at random,* MAR) y proceso no aleatorio (*missing not at random,* MNAR). El primero aparece cuando las unidades con los datos completos son similares a las de datos incompletos; es decir, los sujetos con datos incompletos constituyen una muestra aleatoria simple de todos los sujetos que conforman la muestra. Supongamos que en una encuesta nacional se necesitan estudios costosos, como los electrocardiogramas; podría entonces seleccionarse una submuestra mediante muestreo aleatorio simple de los encuestados, para que se aplique este examen.

En el segundo caso, los sujetos con información completa difieren del resto. Los patrones de los datos ausentes se pueden predecir a partir de la información contenida en otras variables y no de la variable que está incompleta. En el ejemplo anterior, si en lugar de seleccionar la submuestra mediante muestreo simple aleatorio se seleccionara estratificada por subgrupos de edad y género, los valores ausentes para las unidades no incluidas en la submuestra se distribuirían según un mecanismo MAR. En el último caso, el patrón de los datos ausentes no es aleatorio y no se puede predecir a partir de la información contenida en otras variables. En este mecanismo, contrario al MAR, el proceso de ausencia de los datos sólo se explica por los datos que están ausentes (p. ej., un ensayo sobre la pérdida de peso en que un participante abandona el estudio debido a preocupaciones por su pérdida de peso). Para identificar el mecanismo que describe la distribución de los valores ausentes, puede consultarse a Verbeke y Molenberghs[5].

Procedimientos para el manejo de datos incompletos

Se utilizan con frecuencia dos estrategias para el tratamiento de los datos incompletos. La primera, y más

usada, consiste en eliminar las unidades con información incompleta, y la segunda, en imputar los valores ausentes.

Análisis de los casos completos

En este caso se eliminan las unidades que tienen información incompleta y, posteriormente, se realiza el análisis en una base de datos con un número menor de unidades, pero completa en información. Este procedimiento, aunque facilita el análisis por ser muy simple, tiene numerosos problemas que han sido ampliamente discutidos en los últimos años[4,5]; entre ellos, la pérdida de información que implica, que puede producir un impacto espectacular en la precisión y la potencia de las estimaciones. Además, los sesgos pueden ser graves cuando el mecanismo que genera los datos ausentes es MAR y no MCAR. Desafortunadamente, el supuesto MCAR resulta mucho más restrictivo que el MAR y en la práctica tiene lugar con pocas excepciones.

Imputaciones de los valores ausentes

Otra manera de obtener una base de datos en la cual los métodos de análisis para datos completos se puedan utilizar es imputar el valor ausente por otro obtenido, usualmente, de las unidades que contienen la información. Los métodos de imputación son diversos y pueden ser simples, cuando se asigna un único valor, o múltiples, en las ocasiones donde se asignan varios valores al dato ausente.

Entre los métodos de imputación más difundidos se encuentran los siguientes: asignar la media al valor ausente, predecir el valor ausente mediante modelos de regresión e imputaciones múltiples. Para imputar por el primer procedimiento, se calcula el valor promedio de la variable con los casos disponibles y después se asigna este valor a los individuos que no tienen el dato. Éste es uno de los métodos más antiguos y aparece implementado en la mayoría de los paquetes estadísticos.

Cuando se imputa utilizando modelos de regresión, hay que tener en cuenta el tipo de variable que tiene la información incompleta. Si el valor que ha de imputarse es un número (p. ej., la edad, el salario o los valores de presión arterial), se puede emplear la regresión múltiple. En el caso que sea una variable categórica, como el sexo, el estatus socioeconómico o la práctica de ejercicio físico en el tiempo libre, podría emplearse la regresión logística y hacer la imputación según la probabilidad que el modelo de regresión estimado otorgue a cada categoría para el sujeto en cuestión.

Las imputaciones por medias y modelos de regresión son válidas cuando el mecanismo que genera los valores ausentes es MCAR. Estos procedimientos, al igual que el análisis de los casos completos, tienen la ventaja de que se trabaja con una base de datos completa, que se puede analizar empleando los procedimientos y paquetes estadísticos estándares. Sin embargo, la ventaja de estos 2 procedimientos sobre el análisis de casos completos está en el hecho de que no hay pérdida de información, puesto que se trabaja con todas las unidades que fueron estudiadas.

El procedimiento de las imputaciones múltiples[1,2] se refiere a reemplazar cada valor ausente con más de un valor imputado. Es un enfoque basado en simulaciones donde a cada valor ausente se asignan $m > 1$ valores extraídos de una distribución predictiva, lo que produce m bases de datos. Después, en cada base de datos se realiza el análisis estadístico que responda al propósito del estudio, desde obtener estimaciones puntuales y sus intervalos de confianza hasta modelos de regresión. En este caso se obtienen tantos resultados del análisis realizado como imputaciones se hayan hecho. Por ejemplo, si se efectuaron 5 imputaciones y se quería estimar la prevalencia de cierta enfermedad, se tienen 5 estimaciones de dicha prevalencia. Finalmente, se combinan estas estimaciones mediante la regla de Rubin para llegar a la estimación definitiva, que es la que se interpreta como resultado final.

La distribución predictiva se construye a partir de los valores observados; por ejemplo, usualmente se supone que el conjunto de variables sigue una distribución normal multivariada. Para construir la distribución se necesita estimar sus parámetros: vector de medias, y matriz de varianzas y covarianzas. Estas estimaciones se obtienen a partir de las unidades que tienen todos los valores observados. Una vez estimados los parámetros de la distribución, se extraen muestras independientes de ella para asignar los valores en las observaciones que no están completas; según el número de muestras que se seleccione, se tendrá tantas bases de datos para analizar.

Material y métodos

Con la finalidad de ilustrar la aplicación de los procedimientos para el manejo de datos incompletos descritos previamente, se elaboró un ejemplo hipotético con fines didácticos. Cabe aclarar que el ejemplo no se diseñó con la finalidad de comparar el grado de eficiencia con que se desempeña cada procedimiento, sino con la intención exclusiva de ilustrar su aplicación al analizar datos con información incompleta. Supongamos que se quiere estudiar la relación entre la obesidad y la hipertensión arterial y para ello se cuenta con infor-

mación sobre la edad, el sexo, los valores de presión arterial, las mediciones del peso y la talla. Este problema se abordará ajustando los modelos de regresión logística. En todos los casos se modeló la probabilidad de ser hipertenso en función de la edad, el sexo y el índice de masa corporal (IMC), categorizado en bajo peso (IMC < 18,5), normopeso (18,5 ≤ IMC < 25) y sobrepeso (IMC ≥ 25).

Los datos para el ejemplo proceden de una muestra simple aleatoria de 1.000 sujetos de la «Segunda encuesta nacional de factores de riesgo y afecciones no transmisibles», realizada en Cuba durante el año 2001, donde se encuestó a 22.851 personas. La descripción y las medidas de resumen de las variables incluidas en el ejemplo se presentan en la tabla 1.

Para ilustrar los procedimientos antes mencionados se necesitaba un conjunto de datos incompletos. Para ello se decidió que las variables con información incompleta fueran la presión arterial diastólica (PAD), la presión arterial sistólica (PAS), el peso corporal y la talla. Se fijó que la tasa global de no respuesta fuera de un 30%. En las variables PAD y PAS, se eliminó la información contenida en el 20% de las observaciones, y otro tanto en las variables peso y talla, de manera que el 10% de las unidades tuviera información incompleta en las 4 variables (fig. 1). El mecanismo MAR se usó para generar las unidades incompletas, es decir, las observaciones con valores ausentes no constituyen una MSA de la muestra total. Los valores ausentes se generaron con mayor frecuencia en los ancianos que en los jóvenes; de esta manera, la probabilidad de que una unidad tenga valores ausentes se puede predecir por la edad, lo que implica que las unidades con valores ausentes no constituyen una MSA de la muestra original.

Para tratar los datos incompletos se usaron los 4 procedimientos descritos anteriormente:

Casos completos

Se eliminaron las unidades que tuvieran la información incompleta en al menos una variable, lo que dio como resultado una base con 700 observaciones.

Tabla 1. Medidas de resumen de las variables seleccionadas

Variables	Descripción	Media	Error estándar
Edad	Edad (años cumplidos)	48,6	0,58
Sexo	1 masculino, 0 femenino	0,55	0,016
PAD	PAD (mmHg)	80,1	0,39
PAS	PAS (mmHg)	125,6	0,54
Peso	Peso (kg)	65,0	0,45
Talla	Talla (m)	1,60	0,004

PAD: presión arterial diastólica; PAS: presión arterial sistólica.

Figura 1. Datos multivariados con observaciones incompletas simuladas

Unidades	Edad	Sexo	PAS	PAD	Peso	Talla
1	×	×	?	?	×	×
⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	×	×	?	?	×	×
101	×	×	×	×	?	?
⋮	⋮	⋮	⋮	⋮	⋮	⋮
200	×	×	×	×	?	?
201	×	×	?	?	?	?
⋮	⋮	⋮	⋮	⋮	⋮	⋮
300	×	×	?	?	?	?
301	×	×	×	×	×	×
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.000	×	×	×	×	×	×

Nota: El símbolo ? representa los valores ausentes y ×, los observados

Imputaciones por la media

Se calculó el valor promedio de cada variable con las observaciones que tenían información. Posteriormente, se asignó este valor en las unidades que no tenían información en la variable en cuestión y se obtuvo una base con 1.000 observaciones.

Imputaciones mediante modelos de regresión lineal

Se ajustó un modelo para cada variable imputada como variable de respuesta, y la edad y el sexo como variables predictoras. Después se asignó el valor predicho por este modelo a cada una de las observaciones que no tenían información para cada variable analizada, lo que generó también una base con 1.000 observaciones.

Imputaciones múltiples

Debido a que se ha demostrado que de 3 a 5 imputaciones son suficientes para obtener excelentes resultados[1,2], se realizaron 5 para cada valor ausente, lo que generó 5 bases de datos con 1.000 observaciones cada una.

Finalmente, se ajustó un modelo de regresión logística en las 9 bases de datos creadas: la base original, que se usó como referencia; la resultante de aplicar el análisis de casos completos; la obtenida al imputar por la media; la derivada al imputar por modelos de regresión, y las 5 bases obtenidas de las imputaciones múltiples. Por último, se compararon las estimaciones de las *odds ratio* (OR) y sus intervalos de confianza (IC)

obtenidos en cada procedimiento con los calculados en los datos de referencia.

Los procedimientos estadísticos usados, tanto para el manejo de los datos incompletos como para el ajuste de los modelos de regresión logística, se implementaron en el paquete estadístico SAS versión 8.02. Las imputaciones múltiples se realizaron con el procedimiento MI y las estimaciones se combinaron con MIANALIZE.

Resultados

El modelo de regresión logística ajustado con los datos de referencia muestra que, a medida que aumenta la edad, el riesgo de ser hipertenso se incrementa (OR = 1,58), que los hombres tienen el 44% más riesgo de ser hipertensos (OR = 1,44) y que las personas con sobrepeso tienen 2 veces más riesgo de ser hipertensas que aquellas que son clasificadas como normopeso (OR = 2,22) (segunda columna de la tabla 2).

Cuando se usó el análisis de casos completos, las estimaciones de las *odds ratio* del sexo y de la condición de tener sobrepeso fueron las que exhibieron la menor variación, sobreestimada en el 7% y el 2% respectivamente. Sin embargo, la *OR* de la condición de tener bajo peso se subestimó en un 16%. OR de la edad fue lo más afectado, con una subestimación del 65%, llegando incluso a cambiar el sentido de la relación. Con este procedimiento se obtuvieron intervalos de confianza para las *OR* más amplias en comparación con los que se obtuvieron con los datos de referencia, con excepción del estimado para el género (columna de la tabla 2).

Al imputar por las medias y por modelos de regresión, los resultados fueron similares. En ambos las *OR* se subestimaron por encima del 10%. La estimación de la *OR* de la edad fue la que más cambió, con un 61% y un 58% respectivamente, y en ambos se invierte la relación entre la edad y la hipertensión. Los intervalos de confianza son más amplios para las *OR* de la edad y de la condición de bajo peso en ambos procedimientos (cuarta y quinta columnas de la tabla 2).

Cuando se realizó el análisis mediante las imputaciones múltiples, las *OR* se subestimaron y su rango de variación disminuyó considerablemente. Este osciló entre el 1% de *OR* para el sexo y el 16% de *OR* de la condición de tener bajo peso. Hay que resaltar que éste es el único análisis en el cual la relación entre la edad y la hipertensión arterial no se modifica. Los errores de las estimaciones de las *OR* son menores que los obtenidos con los datos de referencia (sexta columna de la tabla 2).

Discusión

Los métodos para el manejo de los datos incompletos aquí ilustrados no son los únicos. Little y Rubin[4] y Verbeke y Molenberghs[5] discuten muchos otros. El método de las imputaciones múltiples, al igual que en otros estudios[6-12], fue el que proveyó las estimaciones más cercanas a las de referencia. Una de las posibles razones asociadas a este resultado es que entre los supuestos de este procedimiento está que el mecanismo que describe la distribución de los valores ausentes es MAR. Con este procedimiento se obtienen razonables estimaciones del valor ausente y una variabilidad alrededor de él con un grado apropiado de incertidumbre. Debido a que con este procedimiento no se predice el valor ausente, sino que se modela la incertidumbre que genera la ausencia de los datos, se preservan las relaciones entre las variables cuando se realizan las imputaciones de los datos ausentes[2]. Los resultados encontrados al realizar este análisis son coherentes con esta afirmación, pues fue este procedimiento el único que preservó la relación entre la edad y la hipertensión arterial.

Tabla 2. *Odds ratio* (OR) e intervalos de confianza para la edad, el sexo y el índice de masa corporal sobre la hipertensión arterial y los procedimientos de análisis

Variables	Datos de referencia	Casos completos	Imputación por medias	Imputación por modelos de regresión	Imputaciones múltiples[a]
Edad[b]	1,58 (1,45-1,72)	0,56 (0,39-0,80)	0,62 (0,46-0,85)	0,66 (0,49-0,90)	1,39 (1,27-1,52)
Sexo					
Varones	1,44 (1,08-1,94)	1,54 (1,38-1,73)	1,23 (1,13-1,33)	1,28 (1,18-1,39)	1,42 (1,06-1,56)
Índice de masa corporal					
Bajo peso	0,49 (0,33-0,77)	0,41 (0,16-0,91)	0,32 (0,13-0,75)	0,42 (0,20-0,95)	0,41 (0,26-0,52)
Sobrepeso	2,22 (1,64-3,03)	2,27 (1,56-3,23)	1,59 (1,14-2,17)	1,69 (1,23-2,27)	1,89 (1,27-2,21)
Normopeso	Referencia	Referencia	Referencia	Referencia	Referencia

[a]La estimación que se muestra es el resultado de combinar las estimaciones de los 5 análisis realizados con este método.
[b]OR para incrementos en 10 años de la edad.

A pesar de que el análisis de los casos completos se implementa fácilmente y está incluido en muchos de los procedimientos de los paquetes estadísticos estándares, debe ser usado con cautela. En este ejemplo, el método de análisis nos condujo a obtener estimaciones muy distantes de las de referencia. Esto pudo deberse a que se presume que los sujetos que tienen información completa se comportan de forma similar a los que, por razones desconocidas, no la tienen. Cabe aclarar que esto generalmente no ocurre así, pues las razones por las cuales los sujetos participantes en el estudio no responden, en la mayoría de las ocasiones, están ligadas a los propósitos del estudio. Además, la pérdida sustancial de información que este método genera reduce el tamaño muestral y, por ende, se verá afectado el proceso de estimación, ya sea en la precisión o en la potencia del estudio. Este problema es más notorio cuando se desea emplear técnicas multivariadas[5].

La similitud en las estimaciones obtenidas por el método de imputación por medias y modelos de regresión resultó semejante a la detectada en otros estudios[6-9]. Las estimaciones obtenidas por ambos métodos subestiman los errores estándares y distorsionan la relación entre la edad y la hipertensión. Esto puede deberse a que los valores ausentes fueron eliminados por un mecanismo MAR dependiente de la edad. Cabe aclarar que estos procedimientos son válidos en el supuesto de que el mecanismo que subyace es MCAR. Además, las diferencias obtenidas en las estimaciones pueden deberse a la pérdida de variabilidad y a la elevada tasa de no respuesta (recuérdese que ésta se fijó en un 30%). Un inconveniente importante de este análisis es que, cuando existen muchos valores ausentes y son sustituidos por la media, se podría producir una homogeneidad artificial que reduce la estimación de los errores. Aunque las imputaciones por modelos de regresión podrían en principio preservar la variabilidad de los datos, no ocurrió así en el ejemplo.

La ventaja principal de los métodos de imputación está en que se dispone de una base de datos completa, que contiene la mayor información posible y permite el uso de los métodos de análisis estándares para datos completos. Sin embargo, se debe tener cuidado con los métodos de imputación que se usen, ya que pueden introducir sesgos mayores que los producidos por la no respuesta.

La elección del procedimiento para el manejo de datos incompletos resulta una tarea compleja, pues un mismo método en determinadas situaciones produce estimaciones precisas y en otras, no. Esto sugiere a los investigadores que, cuando manejen datos incompletos, valoren previamente el uso de más de una alternativa para tratarlos y realicen un análisis de sensibilidad que les permita una mejor elección del procedimiento a implementar.

Bibliografía

1. Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley, 1987.
2. Shaffer JL. Analysis of incomplete multivariate data. London: Chapman and Hall, 1997.
3. Bonet M, Mas P, Chang M, et al. Segunda encuesta de factores de riesgo y afecciones no trasmisibles. La Habana: INHEM-ONE-MINSAP, 2002.
4. Little RJA, Rubin DB. Statistical Analysis with missing data. New York: John Wiley & Sons, 1987.
5. Verbeke G, Molenberghs G. Linear mixed models for longitudinal data. New York: Springer-Verlag, 2000.
6. Crawford SL, Tennstedt SL, McKinlay JB. A comparison of analytic methods for non-random missingness of outcome data. J Clin Epidemiol 1995;48:209-19.
7. Musil CM, Warner CB, Yobas PK, Jones SL. A comparison of imputation techniques for handling missing data. West J Nursing Res 2002;24:815-29.
8. Streiner DL. The case of the missing data: methods of dealing with dropout and other research vagaries. Can J Psychiatry 2002;47:68-75.
9. Hunsberger S, Murray D, Davis CE, Fabsitz RR. Imputation strategies for missing data in a school-based multi-centre study: the Pathway study. Stat Med 2001;20:305-16.
10. Zhou XH, Eckert GJ, Tierney WM. Multiple imputation in public health research. Stat Med2001;20:1541-9.
11. Taylor JM, Cooper KL, Wei JT, Sarma AV, Raghunathan TE, Heeringa SG. Use of multiple imputation to correct for nonresponse bias in a survey of urology symptoms among African-American men. Am J Epidemiol 2002;156: 774-82.
12. Garfield R, Leu CS. A multivariate method for estimating mortality rates among children under 5 years from health and social indicators in Iraq. Int J Epidemiol 2000;29:510-5.