

MISSING YOU

O QUÉ HACER CUANDO NOS FALTAN DATOS...

ME FALTA UN DATO, ¿IMPORTA?

- Depende!!
 - ¿Es en una variable importante?
 - ¿Ocurre a menudo?
 - ¿Puedo deducir/aproximar (con alta confianza) el dato que me falta?
 - ¿Por qué me falta el dato?

ES DE PEROGRULLO, PERO...

- HAGAMOS UN ESFUERZO POR DISPONER DE TODOS LOS DATOS Y CON CALIDAD

¿QUÉ PASA SI ME FALTA UN DATO?

- ERROR ALEATORIO
- SESGO

ME FALTA UN DATO. ¿QUÉ HAGO?

- No hay solución (excepto recuperar el dato)...
- ... pero se puede APROXIMAR (/ESTIMAR /IMPUTAR)
- Esta aproximación no podremos nunca comprobar que sea correcta (sin disponer del dato real)

ME FALTA UN DATO, ¿JUGAMOS?

- Simulaciones (excluyendo datos que conozco y probando posibles soluciones)
- Comparando nuestra muestra con la población (prevalencias, asociaciones, etc)
- Análisis de sensibilidad

ALGUNAS SOLUCIONES

- Excluir la variable que tenga missings
- Análisis de casos completos (CCA)
- Creación de la categoría “missing”
- Imputación de la media (o mediana)
- Imputación del último dato conocido (LVCF) en estudios longitudinales
- Imputaciones más sofisticadas (regresión, etc)
- Imputación múltiple (FCS/MICE, JM/MVN, ...)
- Controlled MI (reference-based, delta-based, ...)
- Fully augmented weighted estimators (FAWEs, AIPW,...)
- ...

¿POR QUÉ ME FALTA EL DATO?

- MCAR: Missing completely at random
- MAR: Missing at random
- MNAR: Missing not at random
- ... block-conditional models, permutation models, block-parallel models,... (via DAG (directed acyclic graphs))
- Missings monótonos y no monótonos (medidas repetidas, seguimiento)
 - Monótonos: a partir de un momento se pierde el dato (censura)
 - No monótonos: el dato aparece y desaparece. Más complejo de tratar.

¿CÓMO INFLUYE LA TA EN LA INCIDENCIA DE ECV?

- Modelo de interés: $I(\text{ECV}) = k + f(\text{TA}) + f(\text{otras variables}) + \text{interacción} + e$
- Lamentablemente, algunos participantes no disponen de una medida de TA
- $R=0$ si el participante no tiene dato de TA, $R=1$ si sí lo tiene
- Y es la TA real (medida o no)
- X son covariables que pueden estar relacionadas con la TA y/o ECV
- $P(R=1 | Y, X)$ [por supuesto, sabiendo $P(R=1)$ sabemos $P(R=0)$]

MCAR

- $P(R=1 | Y, X) = P(R=1)$
- Traducido, la probabilidad de tener el dato (TA) no depende ni del dato (Y) ni de otras variables (X).
- Por ejemplo, se tomó la TA a los pacientes visitados de lunes a viernes en un CAP. Desgraciadamente, el miércoles no funcionaba el tensiómetro y no se pudo medir la TA a un 20% de la muestra.
- Si los pacientes de los miércoles (y sus TAs) son comparables con los del resto de días, tenemos MCAR.

MCAR

- Por ejemplo, se tomó la TA a los pacientes visitados de lunes a viernes en un CAP. Desgraciadamente, el miércoles no funcionaba el tensiómetro y no se pudo medir la TA a un 20% de la muestra.
- La media de las observaciones del miércoles (de TA, ECV y otras variables) debe ser como la del resto de días, OK! Parámetro CCA = parámetro imputado.
- Puedo asignar la TA a los pacientes del miércoles con la de los otros días comparándolos por otras variables (edad, sexo, ...) OK! Parámetro CCA = parámetro imputado. Pero no tengo en cuenta variabilidad! [modelos sofisticados sí la pueden introducir]
- Puedo pensar que estoy en MCAR si la ECV (u otras variables relacionadas) es similar entre los que tienen el dato o no de TA. Aún así, no puedo asegurar MCAR (pero en este ejemplo parece bastante plausible).

MAR

- $P(R=1 | Y, X) = P(R=1 | X)$
- Traducido, la probabilidad de tener el dato (TA) no depende del dato (Y) pero sí de otras variables (X) relacionadas con el dato (TA) y/o el outcome (ECV).
- Por ejemplo, se informó de la TA registrada en el eCAP. La mayoría de menores de 35 años (un 10% de la muestra) no disponían de este dato en sus historias.
- La disponibilidad del dato de TA depende de la edad.
- Es muy probable que las TA registradas en eCAP (pacientes mayores) sean más altas que las TA no registradas (pacientes jóvenes).

MAR

- Por ejemplo, se informó de la TA registrada en el eCAP. La mayoría de menores de 35 años (un 10% de la muestra) no disponían de este dato en sus historias.
- **Parámetro CCA \neq parámetro imputado ya que CCA son más viejos y tienen mayor TA.**
- **Pero puedo imputar un valor a un paciente joven si sé los valores de otros jóvenes (y su perfil de sexo, comorbilidad, etc)**
- Puedo detectar MAR si veo que el % de missings en TA es diferente por edad

MNAR

- $P(R=1 | Y, X) = P(R=1 | Y)$
- Traducido, la probabilidad de tener el dato (TA) depende del dato (Y).
- Por ejemplo, un 10% de la muestra no se presentó a tomarse la TA ya que tenían dolor de cabeza, mareos o vértigo provocados por una elevada tensión arterial [no tenemos porqué conocer este mecanismo...]
- Tener una elevada TA disminuye la probabilidad de ser medida.
- Perdemos a los pacientes con TA más elevadas.

MNAR

- Por ejemplo, un 10% de la muestra no se presentó a tomarse la TA ya que tenían dolor de cabeza, mareos o vértigo provocados por una elevada tensión arterial [no tenemos por qué conocer este mecanismo...]
- **Parámetro CCA \neq parámetro real, pero no lo podemos saber/solucionar.**
- **Este es el peor escenario, sobre todo cuando tenemos missing no monótono.**
Últimamente se están publicando métodos (complicados y poco implementados) para resolverlo (ver Chen, p. ej.).

EJEMPLOS DE TIPOS DE MISSINGS

- MCAR
 - Se pierde una caja con encuestas
 - Se borra accidentalmente un fichero con datos
- MAR
 - Los hombres recuerdan peor cuando vacunaron a sus hijos
 - Los pacientes con cáncer de pulmón recuerdan mejor qué fumaban de jóvenes
- MNAR
 - Los trabajadores con más ingresos son más reservados en declarar cuánto ganan
 - Votantes conservadores no revelan su intención de voto

¿QUÉ PATRÓN SIGUEN MIS MISSINGS?

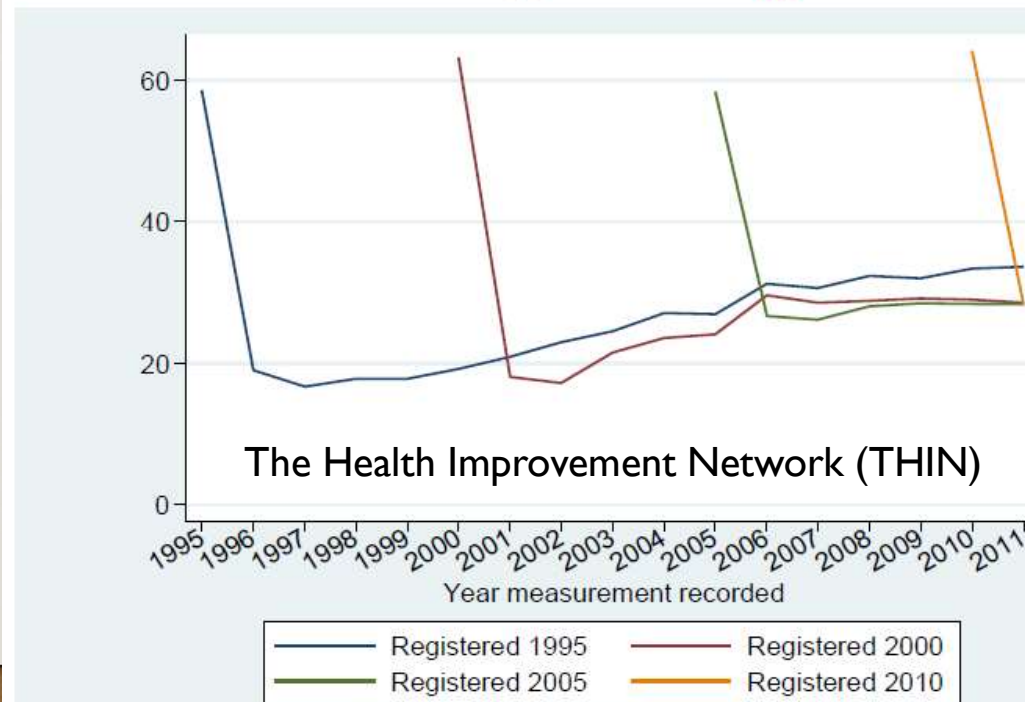
- Es difícil saberlo, y es habitual tener diferentes patrones para diferentes variables con datos ausentes, que incluso pueden interaccionar.

¿QUÉ PATRÓN SIGUEN MIS MISSINGS?



¿Por qué tengo missings en el peso? ¿Existe algún patrón que explique estos missings?

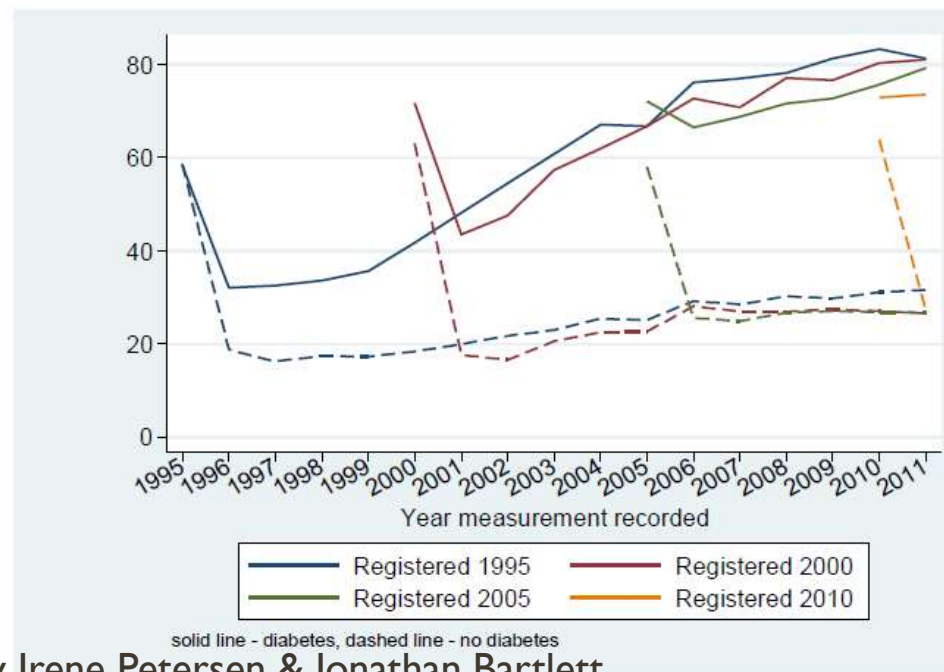
Recording of **weight**



by Irene Petersen & Jonathan Bartlett

¿QUÉ PATRÓN SIGUEN MIS MISSINGS?

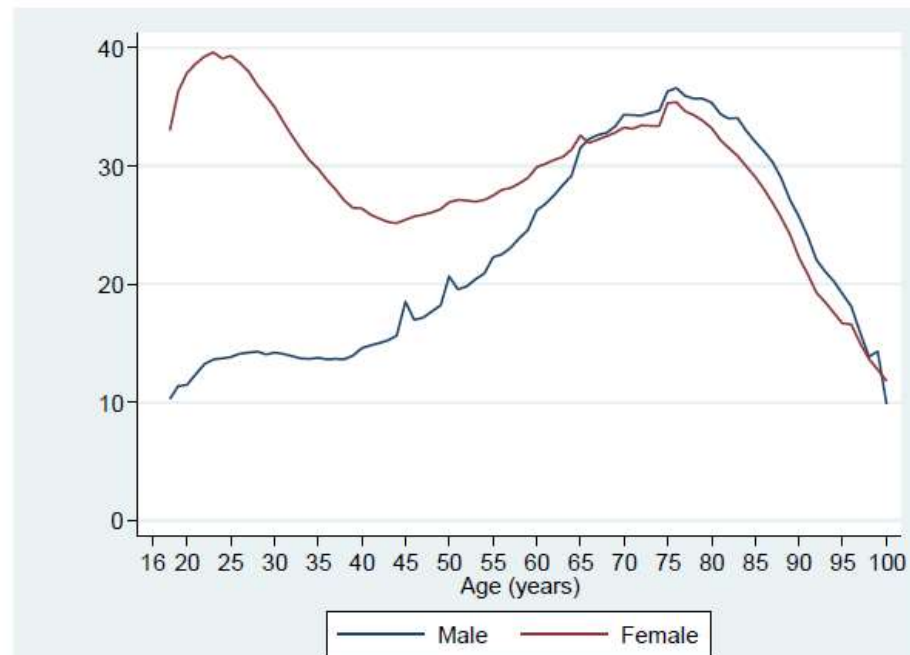
Recording of **weight** in diabetics and non-diabetics



by Irene Petersen & Jonathan Bartlett

¿QUÉ PATRÓN SIGUEN MIS MISSINGS?

Recording of **weight** by age and gender



by Irene Petersen & Jonathan Bartlett

¿CÓMO INFLUYE LA TA EN LA INCIDENCIA DE ECV?

- Modelo de interés: $I(\text{ECV}) = k + f(\text{TA}) + f(\text{otras variables}) + \text{interacción} + e$
- Lamentablemente, algunos participantes no disponen de una medida de TA
- Excluir la variable que tenga missings
- Análisis de casos completos (CCA)
- Creación de la categoría “missing”
- Imputación de la media (o mediana)
- Imputación del último dato conocido (LVCF) en estudios longitudinales
- Imputaciones más sofisticadas (regresión, etc)
- Imputación múltiple (Fully Conditional Specification (FCS) as Multiple Imputation by Chained Equations (MICE), Joint Modelling as Multivariate Normal distribution (MVN), ...)
- Controlled MI (reference-based, delta-based, ...)
- Fully augmented weighted estimators (FAWEs, AIPW,...)
- ...

EXCLUIR LA VARIABLE QUE TENGA MISSINGS

- Inadmisibile si queremos estudiar dicha variable
- Valorar si podría ser un factor de confusión (con lo cual, si la eliminamos introducimos sesgo)

CCA

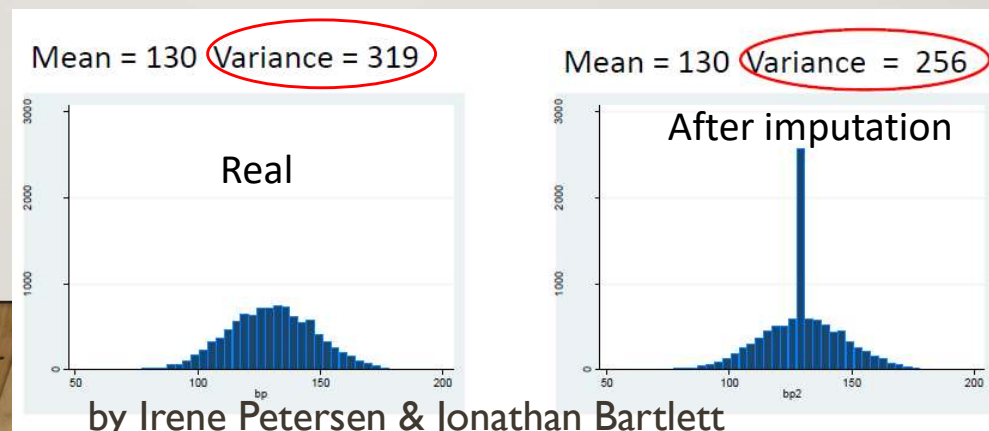
- Pérdida de información existente en otras variables de la observación
- Pérdida de potencia
- Sesgo (excepto si MCAR)
- Muy sencillo
- Muy habitual (seguramente demasiado ;-)

CATEGORÍA MISSING

- Mezcla de diferentes categorías (reales) en una (artificial)
- Sesgo en cualquier dirección (en muchos casos severo)
- No es una solución si quiera para el ajuste
- Potencia (casi) preservada
- En desuso (felizmente ;-)
- Sí puede tener sentido si queremos estudiar por qué hay missings

IMPUTACIÓN DE LA MEDIA (O MEDIANA)

- Sesgo en el valor imputado (excepto MCAR)
- Infraestimación de la varianza (Potencia preservada, e incluso falsamente aumentada)
- No apto para variables categóricas



LVCF

- Asume que los valores son constantes
- Lo que implica una disminución de la varianza (potencia preservada e incluso aumentada)
- Sesgo (especialmente si el efecto existe a corto plazo)
- Se pueden derivar soluciones parecidas (inercia, media del valor anterior y posterior, etc)

IMPUTACIÓN POR REGRESIÓN

- Puede tener en cuenta mucha más información
- Infraestima variabilidad (no tiene en cuenta la incertidumbre que supone la presencia de un missing)
- Posible sesgo (menor que en los anteriores métodos en general)
- Aceptable si el missing es MAR


IMPUTACIÓN MÚLTIPLE (MI)

MI es una técnica estadística flexible, basada en la simulación.

Consta de 3 pasos:

- IMPUTATION: Crea M copias de los datos, reemplazando los missings por imputaciones, basándose en los datos observados (imputation model).
- COMPLETED-DATA ANALYSIS: Analiza nuestro modelo de interés (técnicas estándar) M veces (1 vez para cada dataset (estimation model)).
- POOLING: Combina los M resultados del paso anterior en uno solo.

IMPUTACIÓN MÚLTIPLE (MI)

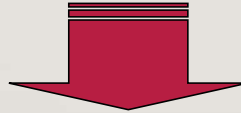
- MI es un procedimiento de simulación
- MI no pretende sustituir los missings en una base de datos, sino tenerlos en cuenta (de una forma “inteligente”) al analizar los datos
- MI está implementado en la mayoría de paquetes () y permite multitud de modelos estadísticos.
- MI tiene en cuenta la incertidumbre provocada por la imputación [IC más anchos].
- Lógicamente, en cada iteración los resultados pueden diferir, ya que también difieren los datasets.

IMPUTACIÓN MÚLTIPLE (MI)

- El modelo infiere correctamente ya que el estimador se basa en los datos observados y en la distribución de los datos imputados dados los observados (MAR).
- No importa si imputamos variables independientes o respuesta.
- No es necesario que el formato de las variables sea idéntico en ambos modelos (por ejemplo, puedo imputar IMC continuo y después usarlo categorizado en el análisis)
- Los modelos de estimación e imputación tienen los mismos problemas que cualquier otro modelaje!!
- En general, el pooling impide algunos procedimientos post-estimation (LRT, GOF, ...) aunque se han estudiado “versiones alternativas” a estos que sí se pueden hacer con MI

IMPUTACIÓN MÚLTIPLE (MI)

- Es necesario incluir, de alguna forma, la relación entre variables dependiente e independientes (estimation model) en los modelos de imputación:
 - Todas las variables que condicionan el hecho de tener missings.
 - Todas las variables que prevemos que usaremos en el modelo de análisis.
 - La estructura de datos que usaremos en el modelo de análisis (pesos, clústers, interacciones,...)



- Si no lo hacemos, tendremos sesgo hacia nulo en los estimadores.
- Este sesgo dependerá del % de missings y de la fuerza de la asociación entre las variables implicadas [poco importante si pocos missings o poca asociación]

IMPUTACIÓN MÚLTIPLE (MI)



- *Omitir la variable respuesta del imputation model: A menudo los missings están en variables explicativas, pero pueden estar relacionadas con la variable respuesta.*
- *Problemas computacionales (?)*
- *MAR: El modelo de imputación debe incluir TODAS las variables del modelo de análisis + las variables que provocan MAR + todas las que se relacionen con la variable con missings.*
- *MNAR: MI puede dar resultados espurios en este caso, con sesgos incluso mayores que CCA.*
- *Se recomienda explorar MI y CCA. Si da resultados muy diferentes, investigar por qué (¿estamos en MNAR?) y publicar ambos resultados.*

IMPUTACIÓN MÚLTIPLE (MI)

TABLE 9 Summary of standalone and MI programmes available in some of the leading statistical software packages

Software for MI: methods derived from multivariate normal			
Data types→	Normal		Mixed response
Data structure→	Independent	Multi-level	Multi-level
Software ↓			
Standalone:	NORM [†]	PAN [†]	REALCOM [*] , PAN [†]
MLwiN	MCMC approach emulates REALCOM		+ 1–2 binary variables
R§	NORM-port	PAN-port	jomo
SAS	PROC MI	–	–
Stata	mi impute mvn	–	–
Software for MI using full conditional specification:			
Software:	functions/packages	Comments	
R§	mi, mice	Available from CRAN; mice based on van Buuren (2018)	
Standalone:	IVEware ⁺	Can be accessed from R, SAS, SPSS, Stata	
SAS	PROC MI	More limited FCS imputation than IVEware	
SPSS	MULTIPLE IMPUTATION	Comes with core package	
Stata	mi impute chained	Comes with the core package	

Key: (†): see Schafer (2001); (*): see Carpenter et al. (2011), uses latent normal model for categorical data; (+): see <https://www.src.isr.umich.edu/software/>; § R has many MI packages; a more complete list is given in the text.

by James R Carter & Melanie Smuk

EJEMPLO REAL (SIMPLIFICADO) FIBROSCAN

¿Cuál es la relación entre IMC y tener una elastografía hepática alterada (Fibroscan ≥ 8 kPa), ajustado por DM2, edad y sexo?

```
. summ f80 imc dm edat sexe
```

Variable	Obs	Mean	Std. dev.	Min	Max
f80	2,712	.0575221	.2328805	0	1
imc	2,712	28.18374	4.835329	16.56065	50.51903
dm	2,712	.1054572	.3071983	0	1
edat	2,712	54.83886	11.63319	19.26078	75.39494
sexe	2,712	.5763274	.4942309	0	1

EJEMPLO REAL (SIMPLIFICADO) FIBROSCAN

¿Cuál es la relación entre IMC y tener una elastografía hepática alterada (Fibroscan ≥ 8 kPa), ajustado por DM2, edad y sexo?

```
. logistic f80 imc dm edat sexe
```

Logistic regression

Log likelihood = -480.09622

Number of obs = 2,712
LR chi2(4) = 233.60
Prob > chi2 = 0.0000
Pseudo R2 = 0.1957

f80	Odds ratio	Std. err.	z	P> z	[95% conf. interval]	
imc	1.215045	.0208709	11.34	0.000	1.174819	1.256647
dm	2.985595	.6043563	5.40	0.000	2.007831	4.439505
edat	1.019168	.0097459	1.99	0.047	1.000245	1.03845
sexe	.3534143	.0671399	-5.48	0.000	.2435438	.5128508
_cons	.0000716	.0000554	-12.34	0.000	.0000157	.000326

Note: **_cons** estimates baseline odds.

EJEMPLO REAL (SIMPLIFICADO) FIBROSCAN

```
. generate aleatori=runiform()
. sort aleatori

. generate imc_mcar=imc if _n>712
(712 missing values generated)

. sort edat

. generate imc_mar=imc if _n>712
(712 missing values generated)

. sort imc

. generate imc_mnar=imc if _n>712
(712 missing values generated)

. summ imc*
```

Generamos 3 variables IMC con distinto tipo de missings
(26% de la muestra):

variable	Obs	Mean	Std. dev.	Min	Max
imc	2,712	28.18374	4.835329	16.56065	50.51903
imc_mcar	2,000	28.18601	4.828749	17.26354	49.04869
imc_mar	2,000	28.88722	4.724833	16.56065	50.43639
imc_mnar	2,000	30.11977	4.042493	24.97704	50.51903

MAR	n	Edad	IMC medio
Sí	712	19-48	26.2
No	2000	48-75	28.9

CCA

EJEMPLO REAL (SIMPLIFICADO) FIBROSCAN

Imputación
de la media

Variable	Obs	Mean	Std. dev.	Min	Max
imc	2,712	28.18374	4.835329	16.56065	50.51903
imc_mcar	2,712	28.18601	4.146448	17.26354	49.04869
imc_mar	2,712	28.88722	4.057215	16.56065	50.43639
imc_mnar	2,712	30.11977	3.47129	24.97704	50.51903

EJEMPLO REAL (SIMPLIFICADO) FIBROSCAN

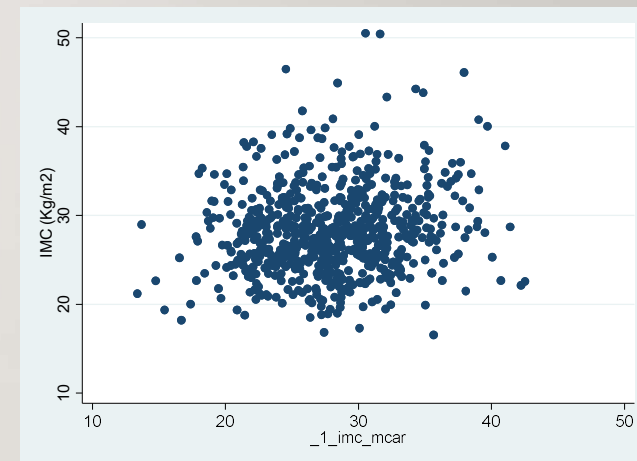
Imputación por
regresión lineal (vs
edad, sexo y AP)

Variable	Obs	Mean	Std. dev.	Min	Max
imc	2,712	28.18374	4.835329	16.56065	50.51903
imc_mcar	2,712	28.19875	4.220216	17.26354	49.04869
imc_mar	2,712	28.56089	4.101007	16.56065	50.43639
imc_mnar	2,712	30.08172	3.48386	24.97704	50.51903

EJEMPLO REAL (SIMPLIFICADO) FIBROSCAN

MI

Variable	Obs	Mean	Std. dev.	Min	Max
imc	2,712	28.18374	4.835329	16.56065	50.51903
imc_mcar (1)	2,712	28.12555	4.814094	13.37204	49.04869
imc_mar (1)	2,712	28.34062	4.754756	14.56536	50.43639
imc_mnar (1)	2,712	29.9586	3.995583	18.56998	50.51903



```
mi impute chained (regress) imc_mcar = f80 dm edat sexe, add(20) rseed(250413)
mi estimate, or: logistic f80 imc_mcar dm edat sexe
```

CCI:

- 0.14 (MCAR)
- 0.10 (MAR)
- 0.00 (MNAR)

EJEMPLO REAL (SIMPLIFICADO) FIBROSCAN

Efecto del IMC sobre $F \geq 8$ kPa								
MCAR	n	OR	SE	IC95%	p	Amplitud IC	OR estimado/real	
Datos reales	2712	1.215	0.021	1.17 1.26	0.000	0.082	1	
CCA	2000	1.214	0.025	1.17 1.26	0.000	0.097	0.999	
Mean imputation	2712	1.202	0.022	1.16 1.25	0.000	0.088	0.990	
Regression imputation	2712	1.205	0.023	1.16 1.25	0.000	0.090	0.992	
MI	2712	1.218	0.023	1.17 1.26	0.000	0.092	1.002	

MAR	n	OR	SE	IC95%	p	Amplitud IC	OR estimado/real	
Datos reales	2712	1.215	0.021	1.17 1.26	0.000	0.082	1	
CCA	2000	1.221	0.024	1.18 1.27	0.000	0.093	1.005	
Mean imputation	2712	1.219	0.023	1.17 1.26	0.000	0.089	1.003	
Regression imputation	2712	1.218	0.022	1.17 1.26	0.000	0.088	1.003	
MI	2712	1.224	0.024	1.18 1.27	0.000	0.093	1.007	

NMAR	n	OR	SE	IC95%	p	Amplitud IC	OR estimado/real	
Datos reales	2712	1.215	0.021	1.17 1.26	0.000	0.082	1	
CCA	2000	1.213	0.023	1.17 1.26	0.000	0.091	0.998	
Mean imputation	2712	1.235	0.024	1.19 1.28	0.000	0.094	1.017	
Regression imputation	2712	1.234	0.024	1.19 1.28	0.000	0.093	1.016	
MI	2712	1.222	0.023	1.18 1.27	0.000	0.091	1.005	

EJEMPLO FICTICIO (SIMPLIFICADO) AQUILGOOD

```
. regress eva tt edat stop sexe
```

Source	SS	df	MS	Number of obs	=	1,300
Model	36675.4581	4	9168.86452	F(4, 1295)	=	673.28
Residual	17635.5278	1,295	13.6181682	Prob > F	=	0.0000
				R-squared	=	0.6753
				Adj R-squared	=	0.6743
Total	54310.9859	1,299	41.8098429	Root MSE	=	3.6903

eva	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
tt	-7.912355	.2332681	-33.92	0.000	-8.36998	-7.45473
edat	.1083265	.0245221	4.42	0.000	.0602192	.1564338
stop	4.487682	.2621724	17.12	0.000	3.973353	5.002011
sexe	.1027309	.2053183	0.50	0.617	-.300062	.5055238
_cons	34.74863	.7263541	47.84	0.000	33.32367	36.17359

Variable	Obs	Mean	Std. dev.	Min	Max
eva	1,300	35.31238	6.466053	16.94459	50.40117
mcar	654	35.30914	6.528623	18.62333	50.40117
mar	654	34.56085	6.719148	16.94459	50.40117
mnar	654	29.8895	3.748495	16.94459	35.94967

MAR	n	Edad	EVA media
Sí	646	29-40	36.1
No	654	18-29	34.6

EJEMPLO FICTICIO (SIMPLIFICADO) AQUILGOOD

Efecto del tt sobre disminución EVA

MCAR	n	β	SE	IC95%	p	Amplitud IC	β estimado/real
Datos reales	1300	-7.9	0.233	-8.4 -7.5	0.000	0.91	1
CCA	654	-8.2	0.328	-8.9 -7.6	0.000	1.29	1.04
Mean imputation	1300	-4.2	0.235	-4.6 -3.7	0.000	0.92	0.53
Regression imputation	1300	-8.2	0.161	-8.6 -7.9	0.000	0.63	1.04
MI	1300	-8.3	0.340	-9.0 -7.6	0.000	1.36	1.05

MAR	n	β	SE	IC95%	p	Amplitud IC	β estimado/real
Datos reales	1300	-7.9	0.233	-8.4 -7.5	0.000	0.91	1
CCA	654	-8.7	0.360	-9.4 -8.0	0.000	1.41	1.09
Mean imputation	1300	-4.5	0.242	-4.9 -4.0	0.000	0.95	0.56
Regression imputation	1300	-8.7	0.168	-9.0 -8.3	0.000	0.66	1.09
MI	1300	-8.6	0.368	-9.4 -7.9	0.000	1.48	1.09

NMAR	n	β	SE	IC95%	p	Amplitud IC	β estimado/real
Datos reales	1300	-7.9	0.233	-8.4 -7.5	0.000	0.91	1
CCA	654	-4.1	0.391	-4.9 -3.3	0.000	1.54	0.52
Mean imputation	1300	-1.0	0.163	-1.3 -0.7	0.000	0.64	0.12
Regression imputation	1300	-4.1	0.146	-4.4 -3.8	0.000	0.57	0.52
MI	1300	-4.1	0.381	-4.9 -3.3	0.000	1.55	0.52

OTROS EJEMPLOS (LITERATURA)

Table 4 Association between BMI and risk of blood transfusion adjusted for age and gender

Patient characteristics	Full data (n=3,500)			Complete case analysis (n=2,733)			Multiple imputation (n=3500, m=5)			Multiple imputation (n=3500, m=30)		
	OR	SE	95% CI	OR	SE	95% CI	OR	SE	95% CI	OR	SE	95% CI
BMI	0.980	0.0085	(0.963, 0.997)	0.978	0.0098	(0.959, 0.997)	0.976	0.0087	(0.959, 0.994)	0.978	0.0098	(0.959, 0.997)
Age (years)												
<75	Baseline											
≥75	2.100	0.1928	(1.754, 2.514)	2.244	0.2421	(1.816, 2.772)	2.097	0.1927	(1.752, 2.511)	2.098	0.1928	(1.752, 2.511)
Gender												
Female	Baseline											
Male	0.815	0.0630	(0.700, 0.948)	0.906	0.0779	(0.765, 1.072)	0.818	0.0633	(0.702, 0.952)	0.817	0.0634	(0.702, 0.951)

Note: Results are presented for full-observed data, complete-case analysis, and multiple imputation and contain point estimates for ORs, SEs, and 95% CIs.

Abbreviations: BMI, body mass index; CI, confidence interval; OR, odds ratio; SE, standard error.

by Alma B Pedersen et al

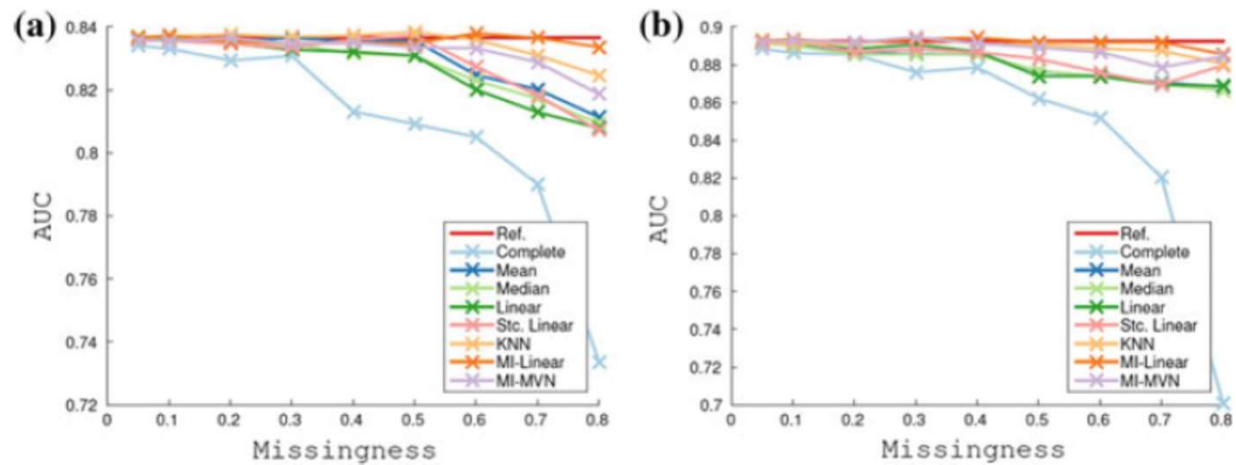


Fig. 13.12 Mean AUC performance of the logistic regression models modelled with different imputation methods for different degrees of univariate missingness of the Age variable

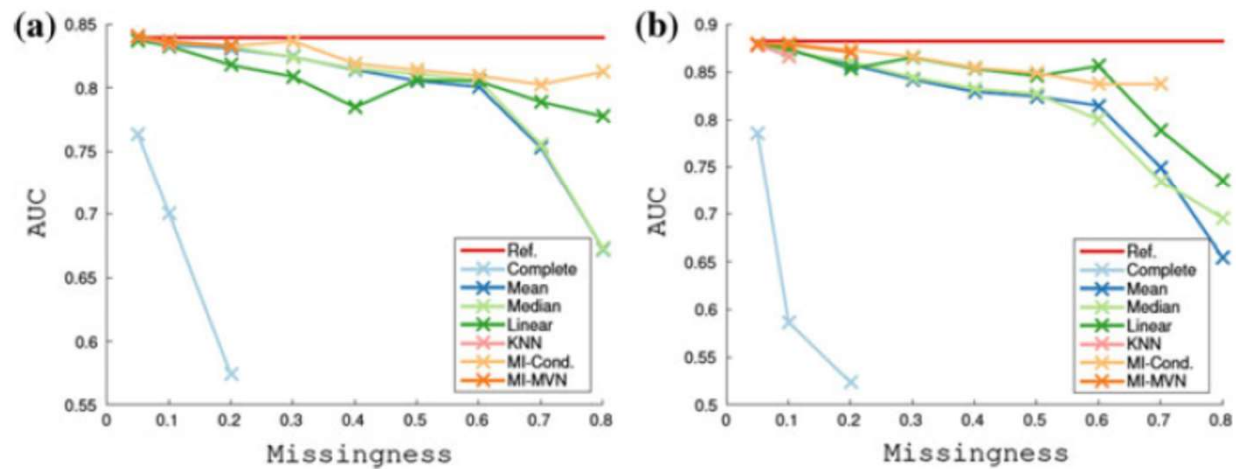


Fig. 13.13 Mean AUC of the logistic regression models for different degrees of multivariate missingness

by Cátia M Salgado et al

¿EXPLICAMOS LA VERDAD?

- *Indicar el % de missings en CADA variable.*
- *Indicar el patrón de missings observado.*
- *Si imputamos, indicar con qué modelos y cuántas iteraciones.*

¿EXPLICAMOS LA VERDAD?

Table 1 | Mean (95% confidence interval) of variables with missing data in complete case dataset (observed values) and imputed dataset, by presence of diabetes

Variables	Missing values (%)	Observed values	Imputed values
Participants without diabetes:			
Total cholesterol (mmol/L)	12 432 (31.9)	5.5 (5.5 to 5.5)	5.4 (5.4 to 5.4)
HDL cholesterol (mmol/L)	17 028 (43.7)	1.6 (1.5 to 1.6)	1.5 (1.5 to 1.6)
LDL cholesterol (mmol/L)	17 138 (44.0)	3.4 (3.4 to 3.4)	3.3 (3.3 to 3.3)
Triglycerides (mmol/L)	15 849 (40.7)	1.3 (1.2 to 1.3)	1.2 (1.2 to 1.2)
Glucose (mmol/L)	12 125 (31.1)	5.3 (5.3 to 5.3)	5.3 (5.3 to 5.3)
Systolic blood pressure (mm Hg)	5118 (13.1)	137.8 (137.6 to 138.0)	137.4 (137.2 to 137.6)
Diastolic blood pressure (mm Hg)	5436 (13.9)	74.8 (74.7 to 75.0)	75.1 (75.0 to 75.2)
Body mass index	10752 (27.6)	28.6 (28.6 to 28.7)	28.1 (28.1 to 28.1)
Participants with diabetes:			
Total cholesterol (mmol/L)	1279 (16.2)	5.2 (5.2 to 5.2)	5.2 (5.1 to 5.2)
HDL cholesterol (mmol/L)	1599 (20.3)	1.4 (1.4 to 1.4)	1.4 (1.4 to 1.4)
LDL cholesterol (mmol/L)	1606 (20.4)	3.2 (3.1 to 3.2)	3.1 (3.1 to 3.1)
Triglycerides (mmol/L)	1478 (18.8)	1.5 (1.5 to 1.5)	1.5 (1.4 to 1.5)
Glucose (mmol/L)	1183 (15.0)	7.8 (7.7 to 7.8)	7.7 (7.7 to 7.8)
Systolic blood pressure (mm Hg)	341 (4.3)	140.6 (140.2 to 141.1)	140.6 (140.2 to 140.9)
Diastolic blood pressure (mm Hg)	408 (5.2)	74.7 (74.4 to 74.9)	74.6 (74.4 to 74.9)
Body mass index	801 (10.2)	29.4 (29.2 to 29.5)	29.2 (29.1 to 29.3)

HDL=high density lipoprotein; LDL=low density lipoprotein.

by Rafa Ramos et al

¿EXPLICAMOS LA VERDAD?

Table 2. Hazard ratios of statin use for incident atrial fibrillation and adverse effects of statins.

	New-users		Non-users		HR (95%CI)
	Events	Incidence rate* (95% CI)	Events	Incidence rate* (95% CI)	
AF, total population	834	10.6 (9.8–11.3)	9039	12.7 (12.5–13.0)	0.91 (0.84–0.99)
AF risk group					
<2.5%	75	3.1 (2.4–3.9)	785	3.6 (3.3–3.9)	0.91 (0.69–1.21)
≥2.5 to <7.5%	420	10.1 (9.1–11.2)	4117	11.6 (11.3–12.0)	0.97 (0.86–1.08)
≥7.5%	338	25.3 (22.5–28.1)	4137	29.6 (28.6–30.5)	0.93 (0.82–1.06)

Table E. AF hazard ratios of statin use and its adverse effects. Complete cases

	New-users		Non-users		HR (95%CI)
	Events	Incidence rate* (95% CI)	Events	Incidence rate* (95% CI)	
AF, total population	385	10.9 (9.8-12.0)	2520	13.0 (12.5-13.5)	0.94 (0.83-1.06)
AF risk groups					
<2.5%	27	2.7 (1.7-3.8)	169	3.4 (2.9-3.9)	0.80 (0.51-1.26)
≥2.5 to <7.5	192	10.0 (8.6-11.4)	1178	11.4 (10.8-12.1)	0.97 (0.82-1.14)
≥7.5	166	26.2 (22.2-30.2)	1173	28.2 (26.6-29.8)	1.04 (0.87-1.24)

by Lia Alves-Cabratosa et al

¿EXPLICAMOS LA VERDAD?

Table 4 Reporting and Handling of Missing Data Issues in the Included Interrupted Time Series Studies (N=60)

	n	(%)
Missing Data – Considered (n=60)		
No	47	(78.3)
Yes	13	(21.7)
Missing Data – % Reported (n=13)		
% Not reported, but declared as an issue to be solved	2	(15.4)
Covariates <30%/outcome <50%	1	(7.7)
Covariates at baseline (<1% each, not combined)	1	(7.7)
Covariates at baseline (<10% each, not combined)	2	(15.4)
Covariates at baseline (<2%, flow chart)	1	(7.7)
Covariates at baseline (<25% each, not combined)	1	(7.7)
Covariates at baseline (<25%, flowchart)	1	(7.7)
Covariates at baseline (<30% each, not combined)	1	(7.7)
Covariates at baseline (<5%, flowchart)	1	(7.7)
Outcome <60%	1	(7.7)
Smoking (one case), outcome irregularly recorded	1	(7.7)
Missing Data Mechanism – Considered (n=13)		
No	11	(84.6)
Yes	2	(15.4)

by Juan Carlos Bazo-Álvarez et al

Missing Data Mechanism – Reported (n=2)		
MAR	1	(50)
MNAR	1	(50)
Method for Handling Missing Data – Considered (n=13)		
No	0	(0)
Yes	13	(100)
Method for Handling Missing Data – Reported (n=13)		
CCA	11	(84.6)
Mixed intercept model for handling missing outcomes	1	(7.7)
Mixed intercept and slope model for handling missing outcomes	1	(7.7)
Sensitivity Analysis for Missing Data Mechanism – Considered (n=13)		
No	11	(84.6)
Yes	2	(15.4)
Sensitivity Analysis for Missing Data Mechanism – Reported (n=2)		
Comparing results from MICE versus CCA	1	(50)
Comparing results from using a “missing data category” versus CCA	1	(50)

Abbreviations: MAR, missing at random; MNAR, missing not at random; CCA, complete case analysis; MICE, multiple imputation by chained equations.

CONCLUYENDO...

- *Reflexionemos por qué tenemos missings.*
- *Valoremos el patrón de missings, démosles la importancia que se merecen.*
- *Las observaciones con missing, ¿siguen algún patrón? Si es sí, seguro que no es MCAR.*
- *Intentar (¿rezar?) que los missings sean MAR (o MCAR)*
- *Tener un estadístico cerca ;-))*

CONCLUYENDO...

- Definir un buen imputation model:
 - Que tiene en cuenta el estimation model
 - Que incluya las variables correlacionadas con los missings
 - Incluir interacciones, efectos no lineales si también las usa el estimation model
- Seleccionar un número suficientemente alto de réplicas (>20?, >50? Stata recomienda 20 mínimo)
- Comparar los resultados con CCA e interpretar
- NO esconder los missings (estamos haciendo CIENCIA)

BIBLIOGRAFIA

- *Alves-Cabratosa L et al. Bazo-Álvarez et al. Statins and new-onset atrial fibrillation in a cohort of patients with hypertension. Analysis of electronic health records, 2006-2015. Plos One 2017. [ejemplo report]*
- *Bazo-Álvarez JC et al. Current Practices in Missing Data Handling for Interrupted Time Series Studies Performed on Individual-Level Data: A Scoping Review in Health Research. Clinical Epidemiology 2021. [Manejo de los missings en la literatura]*
- *van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 2007. [teórico: chained equations vs. joint modelling]*
- *Carpenter JR, Smuk M. Missing data: A statistical framework for practice. Biometrical Journal 2021. [avanzado y actualizado, muchas citas]*

BIBLIOGRAFIA

- *Chen Y. Causal inference and missing data 2019.* http://faculty.washington.edu/yenchic/19A_stat535/Lec12_causal_missing.pdf [breve resumen del problema y de MI y métodos vía DAG. Da algunas opciones (complicadas) para MNAR]
- *Little RJA, Rubin DB. Statistical análisis with missing data. 3rd ed. Hoboken, NJ:Wiley 2020.* [(casi) todo]
- *Marston L et al. Issues in multiple imputation of missing data for large general practice clinical databases. Pharmacoepidemiol Drug Saf 2010.* [ejemplo]
- *Pedersen AB et al. Missing data and multiple imputation in clinical epidemiological research. Clinical Epidemiology 2017.* [MI explicada de forma muy sencilla]

BIBLIOGRAFIA

- *Salgado CM et al. Missing Data. In: Secondary Analysis of Electronic Health Records. MIT Critical Data. Massachusetts Institute of Technology. Cambridge, MA 2016. [básico, fácil de leer]*
- *StataCorp. Stata: Release 17. Statistical Software. College Station, TX: StataCorp LLC 2021. [manual de Stata]*
- *Sterne JAC et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009. [pros y contras de MI]*
- *Ramos R et al. Statins for primary prevention of cardiovascular events and mortality in old and very old adults with and without type 2 diabetes: retrospective cohort study. BMJ 2018. [ejemplo report]*
- *Tan PT et al. A review of the use of controlled multiple imputation in randomised controlled trials with missing outcome data. BMC Medical Research Methodology 2021. [controlled MI]*

MISSING YOU

JOHN WAITE ([HTTPS://WWW.YOUTUBE.COM/WATCH?V=S89AMjFP-j0](https://www.youtube.com/watch?v=S89AMjFP-j0))

Gràcies!!