APPLICAZIONE DELL'IMPUTAZIONE DEI DATI MANCANTI NELLE ANALISI DI SOPRAVVIVENZA SUI DATI DEI REGISTRI TUMORI ITALIANI

Buzzoni C[1], Crocetti E[1], Coviello E[2] , Autelitano M[3], Falcini F[4], Federico M[5], Ferretti S[6], Fusco M[7], Giacomin A[8], Gola G[9], Madeddu A[10], Mazzoleni G[11], Michiara M[12], Pannozzo F[13], Serraino D[14], Tessandori R[15], Traina A[16], Tumino R[17], Vercelli M[18], Zambon P[19] e AIRTUM Working Group[20]

1 Istituto di Prevenzione Oncologica, Firenze;
2 Azienda Sanitaria Locale BT, Barletta,
3 RT Milano,
4 RT Romagna,
5 RT Modena,
6 RT Ferrara,
7 RT Napoli,
8 RT Biella,
9 RT Como,
10 RT Siracusa,
11 RT Alto Adige,
12 RT Parma,
13 RT Latina,
14 RT Fiuli Venezia Giulia,
15 RT Sondrio,
16 RT Palermo mammella,
17 RT Ragusa,
18 RT Genova,
19 RT Veneto,
20 www.registri-tumori.it

Introduzione: Registri Tumori e qualità dei dati

completezza

Confrontabilità

Accuratezza

tempestività

EUROPEAN JOURNAL OF CANCER 43 (2007) 909–913

Interpreting international comparisons of cancer survival: The effects of incomplete registration and the presence of death certificate only cases on survival estimates

David Robinson[a,*], *Risto Sankila*[b], *Timo Hakulinen*[b], *Henrik Møller*[a]

British Journal of Cancer (2005) 92, 576–579

Short Communication

Population-based monitoring of cancer patient survival in situations with imperfect completeness of cancer registration

H Brenner[*,1] and T Hakulinen[2]

Int. J. Cancer: **125,** 432–437 (2009)

Implications of incomplete registration of deaths on long-term survival estimates from population-based cancer registries

Hermann Brenner[1]* and Timo Hakulinen[2]

British Journal of Cancer (2011) 105, 170–176

Completeness of case ascertainment and survival time error in English cancer registries: impact on 1-year survival estimates

H Møller[*,1], S Richards[1], N Hanchett[1], SP Riaz[1], M Lüchtenborg[1], L Holmberg[2] and D Robinson[1]

Introduzione: Registri Tumori e sopravvivenza

Editorials represent the opinions of the authors and not necessarily those of the *BMJ* or BMA

For the full versions of these articles see bmj.com

EDITORIALS

UK cancer survival statistics

Are misleading and make survival look worse than it is

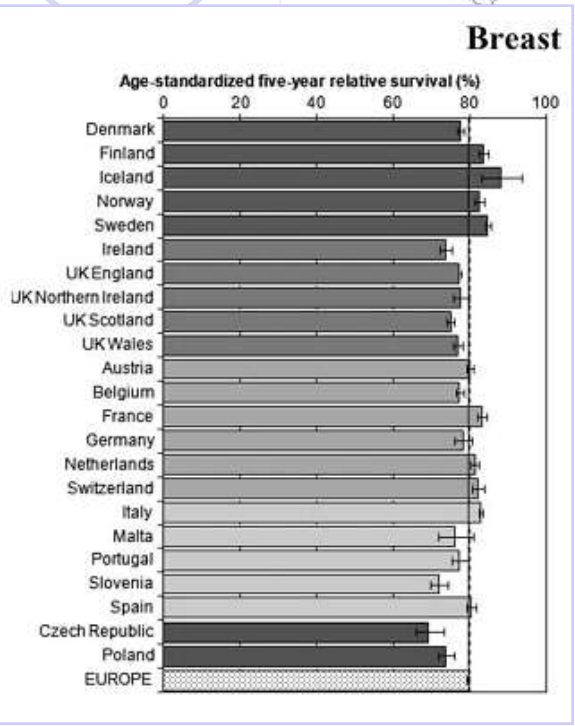
RESEARCH, p 335

Valerie Beral professor of epidemiology, Cancer Epidemiology Unit, University of Oxford, Oxford OX3 7LF
pa.valerie.beral@ceu.ox.ac.uk
Richard Peto professor of medical statistics and epidemiology, Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU), University of Oxford, Oxford OX3 7LF

In the linked article, Autier and colleagues report that (population based) breast cancer mortality rates have fallen over the past two decades in many European countries, with a greater decline in the United Kingdom than in any other large country.[1] That the UK is leading Europe in the speed with which national breast cancer mortality rates are falling is in stark contrast to, and at first sight difficult to reconcile with, claims that survival after breast cancer onset is worse in the UK than elsewhere in western Europe.[2]

The unpromising UK cancer survival estimates are, however, based on registry data make UK cancer survival rates seem significantly worse than they really are.

Information in cancer registries on deaths from cancer is virtually complete because every death certificate that mentions cancer is automatically sent to one of the regional registries that, between them, cover the UK. That cancer is then registered, and further information is sought (not always successfully) from medical records. Death certificates have for decades played an important role in the way UK registries identify people with cancer. Without this source of informa-



BMJ

RESEARCH

Evidence against the proposition that "UK cancer survival statistics are misleading": simulation study with National Cancer Registry data

Laura M Woods, lecturer in cancer epidemiology,[1] Michel P Coleman, professor of epidemiology and vital statistics,[1] Gill Lawrence, director,[2] Jem Rashbass, director,[3] Franco Berrino, director,[4] Bernard Rachet, senior

Coleman MP, Rachet B, Woods L, Berrino F, Butler J, Capocaccia R, Dickman P, Gavin A, Giorgi R, Hamilton W, et al. BMJ. 2011. Rebuttal to editorial saying cancer survival statistics are misleading.
- UK cancer survival statistics. Survival is multifactorial. BMJ. 2010
- UK cancer survival statistics. Survival is multifactorial.Brewster DH, Black RJ. BMJ. 2010
- UK cancer survival statistics. Reflect NHS clinical realities. BMJ. 2010

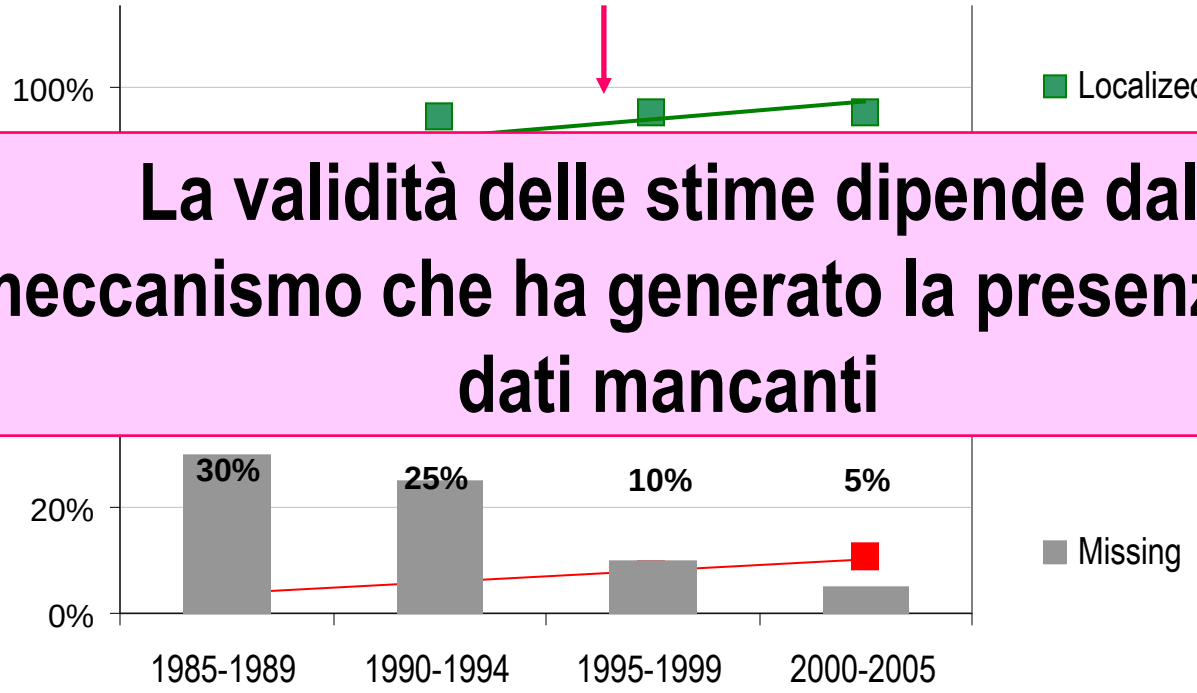
Introduzione

Lo stadio alla diagnosi è un **forte predittore** della sopravvivenza per

valutare l'effetto dei dati mancanti relativi alla definizione dello stadio patologico sulle stime di sopravvivenza di pazienti affette da tumore della mammella presenti nella Banca Dati AIRTUM e valutare l'impatto sulla stima di sopravvivenza dell'imputazione dei dati mancanti mediante il metodo delle imputazioni multiple

Background: Completezza e accuratezza

Registro Tumori regione Toscana: sopravvivenza relativa a 5 anni



La validità delle stime dipende dal meccanismo che ha generato la presenza dei dati mancanti

Definizioni (Rubin)

Missing Completely At Random (MCAR)

P(*x = missing*)

Non dipende dal valore ignoto

Non dipende dai valori osservati nelle altre variabili

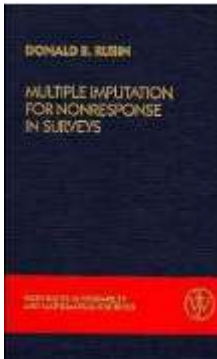
Stime non distorte perdita di efficienza (errori standard ↑)

HOSPITAL	PATIENT-ID	MORPHOLOGICAL-CODE
001	1	8000.3
001	2	8140.3
BUG	X	BUGX. X
002	4	8000.2
BUG	X	BUGX. X
003	6	8000.3
003	7	8200.3
BUG	X	BUGX. X
006	15	8000.2
007	16	8140.3
008	17	8000.3
BUG	X	BUGX. X
009	19	8500.3
009	19	8500.3

Analisi di sopravvivenza

Esclusione dei casi con valori mancanti

Stime non distorte, perdita di potenza



Definizioni (Rubin)

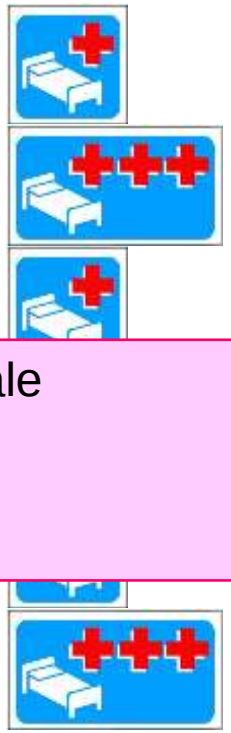
Missing At Random (MAR)

P(*x = missing*)

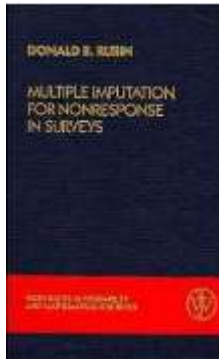
Non dipende dal valore ignoto condizionatamente ai valori osservati nelle altre variabili.

Includere la variabile nel modello

HOSPITAL	PATIENT-ID	MORPHOLOGICAL-CODE
001	1	8000.3
001	2	8140.3
BUG	X	BUGX. X
BUG	X	BUGX. X
BUG	X	BUGX. X
BUG	X	BUGX. X
003	7	8200.3
004	8	8000.2
006	15	8000.2
007	16	8140.3
008	17	8000.3
009	18	8200.3
009	19	8500.3
009	19	8500.3



Analisi di sopravvivenza stratificate per ospedale
Esclusione dei casi con valori mancanti
Stime non distorte, perdita di potenza

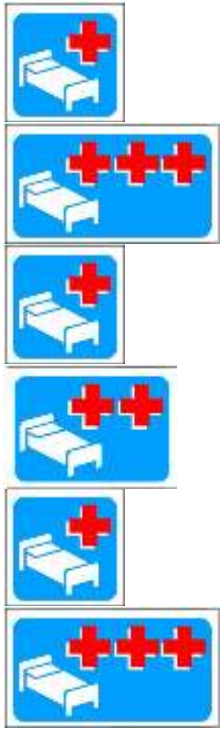


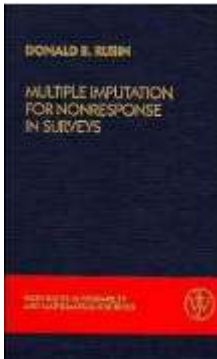
Definizioni (Rubin)

Missing Not At Random (MNAR)

P(*x = missing*)
dipende dal valore ignoto
condizionatamente ai valori osservati nelle altre variabili.

HOSPITAL	PATIENT-ID	MORPHOLOGICAL-CODE
001	1	8000.3
001	2	8140.3
002	3	8500.3
BUG	X	BUGX. X
003	5	8140.3
003	6	8000.3
003	7	8200.3
BUG	X	BUGX. X
004	9	8140.3
004	10	8000.3
004	11	8200.3
005	12	8500.3
005	13	8000.3
005	14	8200.3
BUG	X	BUGX. X
008	17	8000.3
009	18	8200.3
009	19	8500.3
009	19	8500.3





Metodi ad-hoc: imputazioni multiple

- ✓ **Data-set originale**

id	failure	time	stage
1	0	12	early
2	1	7	early
3	1	9	??
4	0	12	advanced
...			
999	1	9	??

- ✓ **M data-set 'completi'**

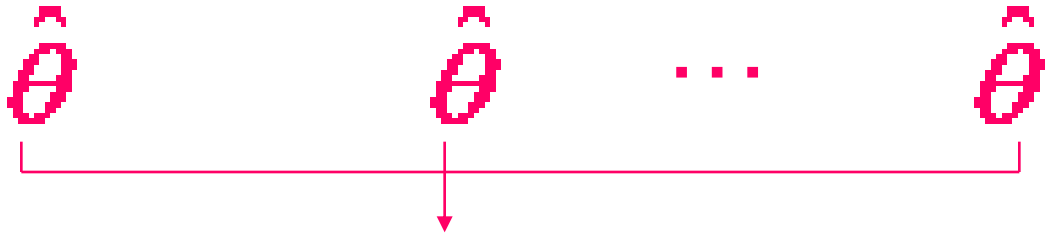
id	failure	time	stage
1	0	12	early
2	1	7	early
3	1	9	early
4	0	12	advanced
...			
999	1	9	advanced

id	failure	time	stage
1	0	12	early
2	1	7	early
3	1	9	advanced
4	0	12	advanced
...			
999	1	9	advanced

. . .

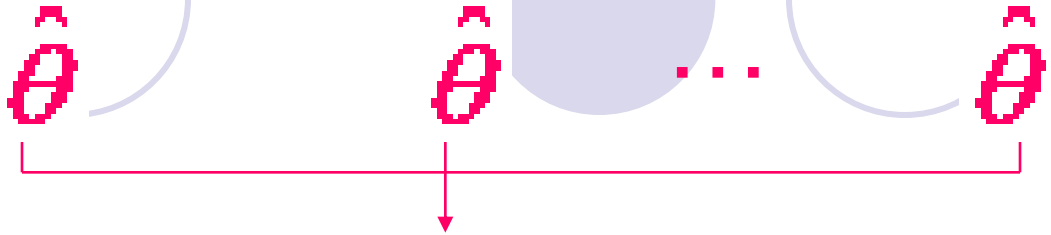
id	failure	time	stage
1	0	12	early
2	1	7	early
3	1	9	advanced
4	0	12	early
...			
999	1	9	advanced

- ✓ **Stima dei parametri di interesse**



Rubin, DB, Multiple Imputation for Nonresponse in Surveys, 1987, New York, Wiley

Metodi ad-hoc per il trattamento dei dati mancanti



Combinazione di

$$\hat{\theta} = \frac{1}{m}\sum_{j=1}^{m}\hat{\theta}_j$$

✓ Stime puntuali

✓ Errori standard

$$\mathrm{var}(\hat{\theta}) = \mathrm{var}_{within}(\hat{\theta}) + \left(\frac{m+1}{m}\right)\mathrm{var}_{beetween}(\hat{\theta})$$

$$\mathrm{var}_{within}(\hat{\theta}) = \frac{1}{m}\sum_{j=1}^{m} W_j$$

$$\mathrm{var}_{beetween}(\hat{\theta}) = \left(\frac{1}{m-1}\right)\sum_{j=1}^{m}\left(\hat{\theta}_j - \hat{\theta}\right)^2$$

ISTITUTO PER LO STUDIO E LA PREVENZIONE ONCOLOGICA

Rubin, DB, Multiple Imputation for Nonresponse in Surveys, 1987, New York, Wiley

Materiali e metodi

DATA:

✓ 18 Registri Tumori (6.5 milioni di donne, 22% della popolazione femminile italiana) con informazione sullo stadio alla diagnosi disponibile nella Banca Dati AIRTUM.

✓ Più di 100.000 della mammella femminile diagnosticati tra il 1978 e il 2006.

✓ Tumori invasivi (ICD-10 C50.*). Follow-up aggiornato al 31.12.2009.

✓ classificazione TNM (AJCC, 2003), categorie I, IIA, IIB IIIA, IIIB, IIIC, IVB

.

Materiali e metodi

Geographic area and Cancer Registries, number and percentage		period					
		<1990	1990-94	1995-99	2000-04	2005-06	Total
AIRTUM pool	I	530 (10%)	3463 (28%)	9612 (31%)	16771 (36%)	6816 (38%)	37192 (33%)
	IIA	718 (13%)	2581 (21%)	6229 (20%)	9757 (21%)	3347 (19%)	22632 (20%)
	IIB	495 (9%)	1404 (11%)	2898 (9%)	4630 (10%)	1423 (8%)	10850 (10%)
	IIIA	179 (3%)	386 (3%)	1236 (4%)	2404 (5%)	1143 (6%)	5348 (5%)
	IIIB	254 (5%)	585 (5%)	1400 (4%)	1754 (4%)	428 (2%)	4421 (4%)
	IIIC	64 (1%)	117 (1%)	224 (1%)	806 (2%)	616 (3%)	1827 (2%)
	IVB	101 (2%)	249 (2%)	912 (3%)	1733 (4%)	671 (4%)	3666 (3%)
	missing	3044 (57%)	3655 (29%)	8768 (28%)	9304 (20%)	3282 (19%)	28053 (25%)
	Total	*5385 (100%)*	*12440 (100%)*	*31279 (100%)*	*47159 (100%)*	*17726 (100%)*	*113989 (100%)*

Materiali e metodi

ANALISI STATISTICHE

Sopravvienza relativa a 5 anni. Stima degli excess hazard ratio (EHR) utilizzando i modelli con spline proposti da Royston (Royston, 2001).

Le analisi di sopravvienza sono state condotte per ogni area geografica.
Variabili esplicative:

- tempo di follow-up [<6 mesi (ref.), 6-11, 12-23, 24-35, 36-4, 48-59 mesi];
- periodo di diagnosi [1995-99 (ref), 2000-04, 2005-06];
- età alla diagnosi[<50 anni (ref.), 50-59, 60-69, 70+ anni];
- gruppo morfologico [Duttale (ref.), Lobulare, Misto (duttale/lobulare), Altri e non spec.];
- registro tumori
- stadio alla diagnosi [I (ref.), IIA, IIB, IIIA, IIIB, IIIC, IVB]
- grading.

stadio alla diagnosi e grading -> valori mancanti

- **complete case analysis**
- **missing category method**
- **multiple imputation**

Materiali e metodi

ANALISI STATISTICHE: procedura di imputazione

do m times

do iteratively until the estimates are stable

Missing value in each variable are substituted with observed value (random draw)

do for each variable with missing value

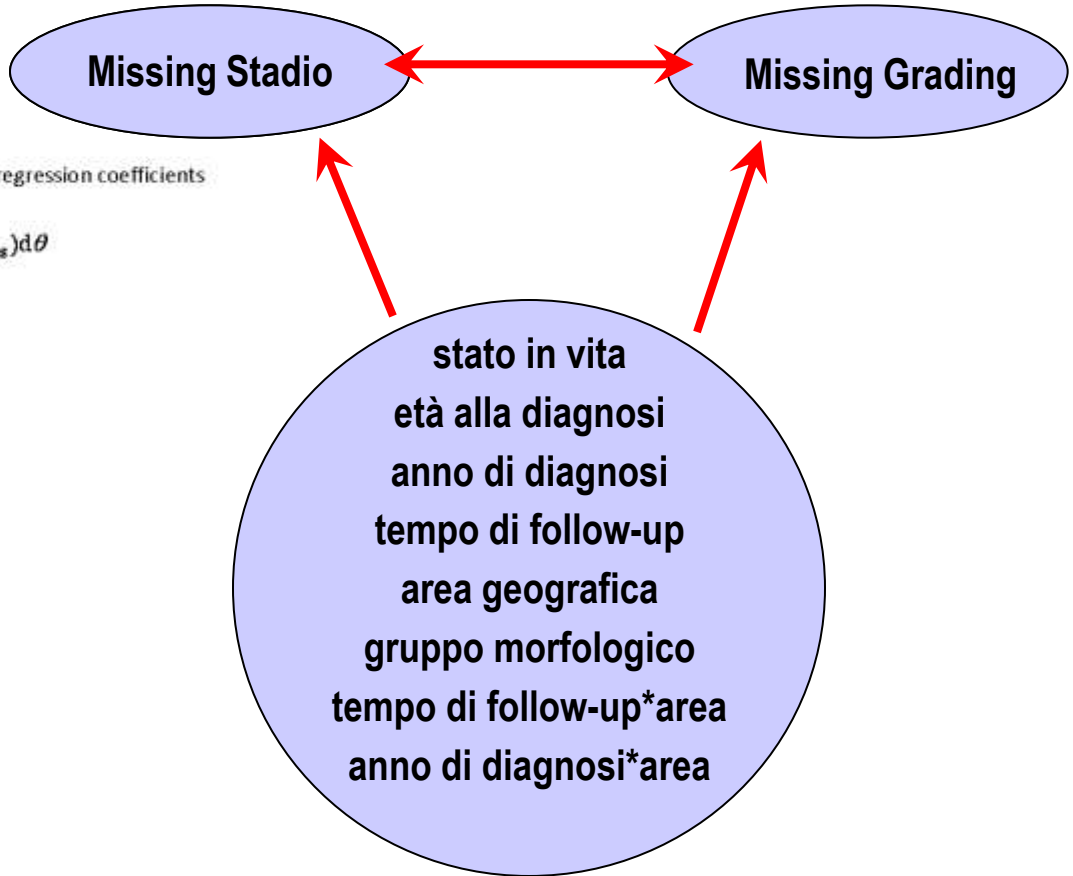
- ✓ Temporarily set aside the value entered
- ✓ Specify the regression model $x_1 = \beta^*(x_2, \ldots, x_k)^T$
- ✓ *Draw* θ^* from the posterior distribution of the regression coefficients
- ✓ Predict x_1 using $\beta^*(x2, \ldots, xk)^T$

end

$$f(\mathbf{Y}_{miss} \mid \mathbf{Y}_{obs}) = \int_{\Theta} f(\mathbf{Y}_{miss} \mid \mathbf{Y}_{obs}, \theta) f(\theta \mid \mathbf{Y}_{obs}) d\theta$$

end

end



Risultati

Var.		complete case analysis				missing category method				multiple imputation			
		EHR	Err. Std.	[95% CI]		EHR	Err. Std.	[95% CI]		EHR	Err. Std.	[95% CI]	
Period of diagnosis	<1990 (ref)	1.0				1.0				1.0			
	1990-94	0.95	0.08	0.81	1.12	0.95	0.08	0.79	1.13	1.07	0.26	0.68	1.71
	1995-99	0.81	0.05	0.71	0.92	0.84	0.06	0.73	0.95	0.62	0.11	0.44	0.88
	2000-04	0.68	0.06	0.57	0.80	0.77	0.06	0.66	0.89	0.42	0.08	0.29	0.61
	2005+	0.56	0.06	0.46	0.69	0.66	0.05	0.56	0.77	0.26	0.03	0.20	0.34
Age groups	<50 years	1.00				1.00				1.00			
	50-59 years	1.08	0.05	0.98	1.20	1.13	0.04	1.05	1.22	1.13	0.05	1.03	1.23
	60-69 years	1.11	0.06	1.00	1.23	1.25	0.06	1.15	1.37	1.33	0.06	1.21	1.46
	70+ years	1.36	0.06	1.25	1.48	1.72	0.08	1.56	1.88	1.86	0.06	1.74	1.99
at diagnosis	IIIA	16.45	1.72	13.40	20.19	17.51	1.99	14.01	21.89	6.50	1.01	4.81	8.78
	IIIB	20.36	2.25	16.40	25.28	20.70	2.38	16.53	25.93	8.30	1.08	6.44	10.69
	IIIC	26.92	2.65	22.20	32.65	27.91	3.16	22.36	34.84	10.67	2.47	6.87	16.57
	IV	94.23	10.50	75.74	117.23	86.84	10.36	68.73	109.73	21.49	4.15	14.84	31.12
	missing					15.23	1.60	12.39	18.71				
Grading	I(ref.)	1.00				1.00				1.00			
	II	2.73	0.28	2.23	3.33	2.63	0.22	2.23	3.10	1.27	0.11	1.08	1.50
	III	5.47	0.68	4.29	6.99	5.46	0.45	4.64	6.42	1.79	0.16	1.50	2.13
	IV					4.33	0.35	3.69	5.08				

Var.		complete case analysis				missing category method				multiple imputation			
		EHR	Err. Std.	[95% CI]		EHR	Err. Std.	[95% CI]		EHR	Err. Std.	[95% CI]	
Age groups	<50 years	1.00				1.00				1.00			
	50-59 years	1.08	0.05	0.98	1.20	1.13	0.04	1.05	1.22	1.13	0.05	1.03	1.23
	60-69 years	1.11	0.06	1.00	1.23	1.25	0.06	1.15	1.37	1.33	0.06	1.21	1.46
	70+ years	1.36	0.06	1.25	1.48	1.72	0.08	1.56	1.88	1.86	0.06	1.74	1.99

AIRTUM-pool and geographic areas. Five-years relative survival, Excess-hazard ratio for stage (reference stage = 1), adjusted by follow-up time, period of diagnosis, age group, cancer registry on *(a)* the cases with complete information for stage variable, *(b)* all cases, considering missing in stage variable as a valid category - missing category method, *(c)* on all cases, after applying multiple imputation

Risultati

Var.		complete case analysis				missing category method				multiple imputation			
		EHR	Err. Std.	[95% CI]		EHR	Err. Std.	[95% CI]		EHR	Err. Std.	[95% CI]	
North-east	I (ref.)	1.00				1.00				1.00			
	IIA	4.56	0.50	3.68	5.66	4.79	0.55	3.82	6.01	2.11	0.11	1.91	2.32
	IIB	9.98	1.09	8.05	12.37	10.36	1.19	8.27	12.98	3.14	0.18	2.81	3.50
	IIIA	16.76	1.84	13.51	20.80	17.23	1.98	13.75	21.59	3.72	0.21	3.33	4.15
	IIIB	21.83	2.44	17.53	27.19	22.09	2.58	17.58	27.77	4.91	0.29	4.38	5.52
	IIIC	28.44	3.26	22.72	35.59	28.85	3.44	22.83	36.44	5.59	0.45	4.77	6.55
	IV	115.44	11.98	94.19	141.48	105.37	11.48	85.11	130.46	9.98	0.64	8.81	11.30
	missing					16.03	1.74	12.96	19.84				
North-west	I (ref.)	1.00				1.00				1.00			
	IIA	5.15	0.91	3.64	7.28	5.86	1.17	3.96	8.68	3.12	0.33	2.54	3.82
	IIB	12.30	2.13	8.75	17.27	14.23	2.80	9.68	20.92	6.53	0.67	5.34	7.97
	IIIA	20.48	3.63	14.46	28.99	23.11	4.63	15.61	34.22	11.31	1.17	9.24	13.84
	IIIB	25.94	4.53	18.42	36.54	28.59	5.64	19.43	42.09	12.32	1.29	10.04	15.12
	IIIC	39.19	7.87	26.44	58.10	38.95	8.53	25.36	59.83	21.26	2.58	16.79	26.91
	IV	99.39	17.30	70.66	139.79	97.04	19.01	66.09	142.47	34.20	3.42	28.15	41.56
	missing					21.63	4.16	14.84	31.54				
Centre	I (ref.)	1.00				1.00				1.00			
	IIA	3.40	0.53	2.50	4.61	3.71	0.62	2.67	5.16	2.13	0.22	1.73	2.61
	IIB	9.26	1.38	6.92	12.40	10.21	1.64	7.45	13.99	3.65	0.38	2.98	4.47
	IIIA	19.76	3.10	14.53	26.87	20.94	3.52	15.07	29.11	6.15	0.78	4.80	7.87
	IIIB	12.65	2.09	9.16	17.47	13.70	2.39	9.73	19.30	5.78	0.71	4.55	7.34
	IIIC	30.40	5.93	20.73	44.57	29.09	5.87	19.59	43.20	10.05	1.77	7.16	14.09
	IV	89.33	13.56	66.35	120.27	70.69	11.38	51.57	96.91	18.02	2.29	14.08	23.07
	missing					11.43	1.76	8.45	15.46				
South	I (ref.)	1.00				1.00				1.00			
	IIA	3.01	0.64	1.98	4.55	3.06	0.66	2.00	4.68	2.37	0.42	1.69	3.32
	IIB	7.18	1.47	4.81	10.71	7.21	1.50	4.79	10.86	4.84	0.76	3.58	6.54
	IIIA	8.83	2.04	5.60	13.90	8.78	2.06	5.54	13.91	7.88	1.49	5.48	11.32
	IIIB	12.51	2.68	8.22	19.03	12.21	2.67	7.96	18.74	9.29	1.59	6.68	12.93
	IIIC	14.65	3.83	8.78	24.46	14.15	3.73	8.44	23.72	13.95	2.78	9.53	20.42
	IV	48.25	9.56	32.72	71.13	46.68	9.45	31.39	69.40	28.27	4.44	20.87	38.28
	missing					9.34	1.92	6.25	13.98				

AIRTUM-pool and geographic areas. Five-years relative survival, Excess-hazard ratio for stage (reference stage = I), adjusted by follow-up time, period of diagnosis, age group, cancer registry on *(a)* the cases with complete information for stage variable, *(b)* all cases, considering missing in stage variable as a valid category - missing category method, *(c)* on all cases, after applying multiple imputation

Risultati

	0-49		50-59		60-69		70+	
	n. and % of analysed cases	% of analysed cases after imputation procedure	n. and % of analysed cases	% of analysed cases after imputation procedure	n. and % of analysed cases	% of analysed cases after imputation procedure	n. and % of analysed cases	% of analysed cases after imputation procedure
I	7972 (34%)	41%	9549 (39%)	46%	11373 (40%)	47%	8298 (22%)	30%
IIA	5318 (23%)	28%	5146 (21%)	25%	5789 (20%)	24%	6379 (17%)	24%
IIB	2692 (12%)	14%	2275 (9%)	12%	2569 (9%)	11%	3314 (9%)	14%
IIIA	1382 (6%)	8%	1218 (5%)	6%	1269 (4%)	6%	1479 (4%)	8%
IIIB	533 (2%)	3%	672 (3%)	4%	1016 (4%)	5%	2200 (6%)	10%
IIIC	377 (2%)	2%	421 (2%)	2%	440 (2%)	2%	589 (2%)	4%
IV	525 (2%)	4%	694 (3%)	4%	879 (3%)	5%	1568 (4%)	11%
missing	4547 (19%)		4383 (18%)		5126 (18%)		13997 (37%)	
	23346 (100%	100%	24358 (100%	100%	28461 (100%	100%	37824 (100%	100%

Distribution of analysed cases by geographic area, TNM stage and period, n. and percentage. Percentage after imputation procedure.
a = n. and % of analysed cases, b = % of analysed cases after imputation procedure

ISTITUTO PER LO STUDIO E LA PREVENZIONE ONCOLOGICA

Conclusioni

✓ Dati di popolazione e stadio del tumore della mammella femminile: l'ipotesi MAR -> buona prima approssimazione

✓ Presentare entrambe le analisi (con e senza dati mancanti) devono essere presentati.

✓ Uso dello stesso metodo.

✓ Imputazione multipla: tecnica appropriata per trattare informazioni mancanti sullo stadio del tumore nei registri tumori di popolazione, se la quantità di casi non stadiati è ad un livello ragionevole.

✓ Stime simili degli EHR per la variabile incompleta stadio alla diagnosi e per le altre variabili: queste analisi rassicurano sulla validità delle valutazioni di sopravvivenza tradizionali svolte sulla base dei dati dei registri dei tumori italiani.

American Journal of Epidemiology
Published by the Johns Hopkins Bloomberg School of Public Health 2008.

Vol. 168, No. 4
DOI: 10.1093/aje/kwn071
Advance Access publication June 30, 2008

Special Article

Use of Multiple Imputation in the Epidemiologic Literature

Mark A. Klebanoff[1] and Stephen R. Cole[2]

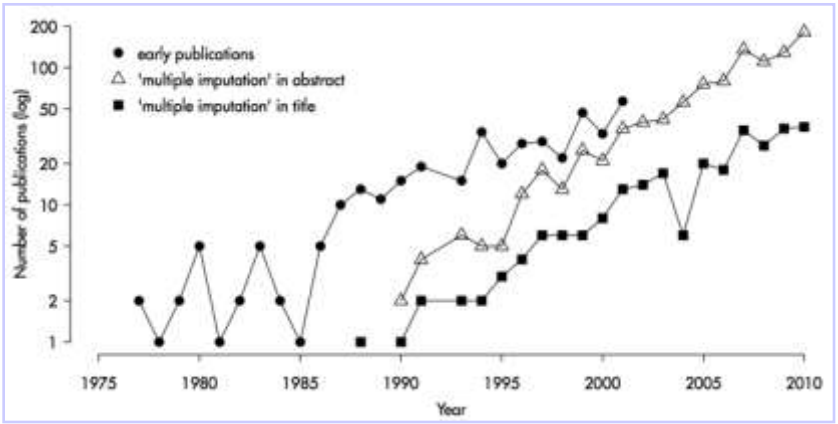
[1] Division of Epidemiology, Statistics, and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Department of Health and Human Services, Bethesda, MD.
[2] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD.

Received for publication January 14, 2008; accepted for publication March 4, 2008.

The authors attempted to catalog the use of procedures to impute missing data in the epidemiologic literature and to determine the degree to which imputed results differed in practice from unimputed results. The full text of articles published in 2005 and 2006 in four leading epidemiologic journals was searched for the text *imput*. Sixteen articles utilizing multiple imputation, inverse probability weighting, or the expectation-maximization algorithm to impute missing data were found. The small number of relevant manuscripts and diversity of detail provided precluded systematic analysis of the use of imputation procedures. To form a bridge between current and future practice, the authors suggest details that should be included in articles that utilize these procedures.

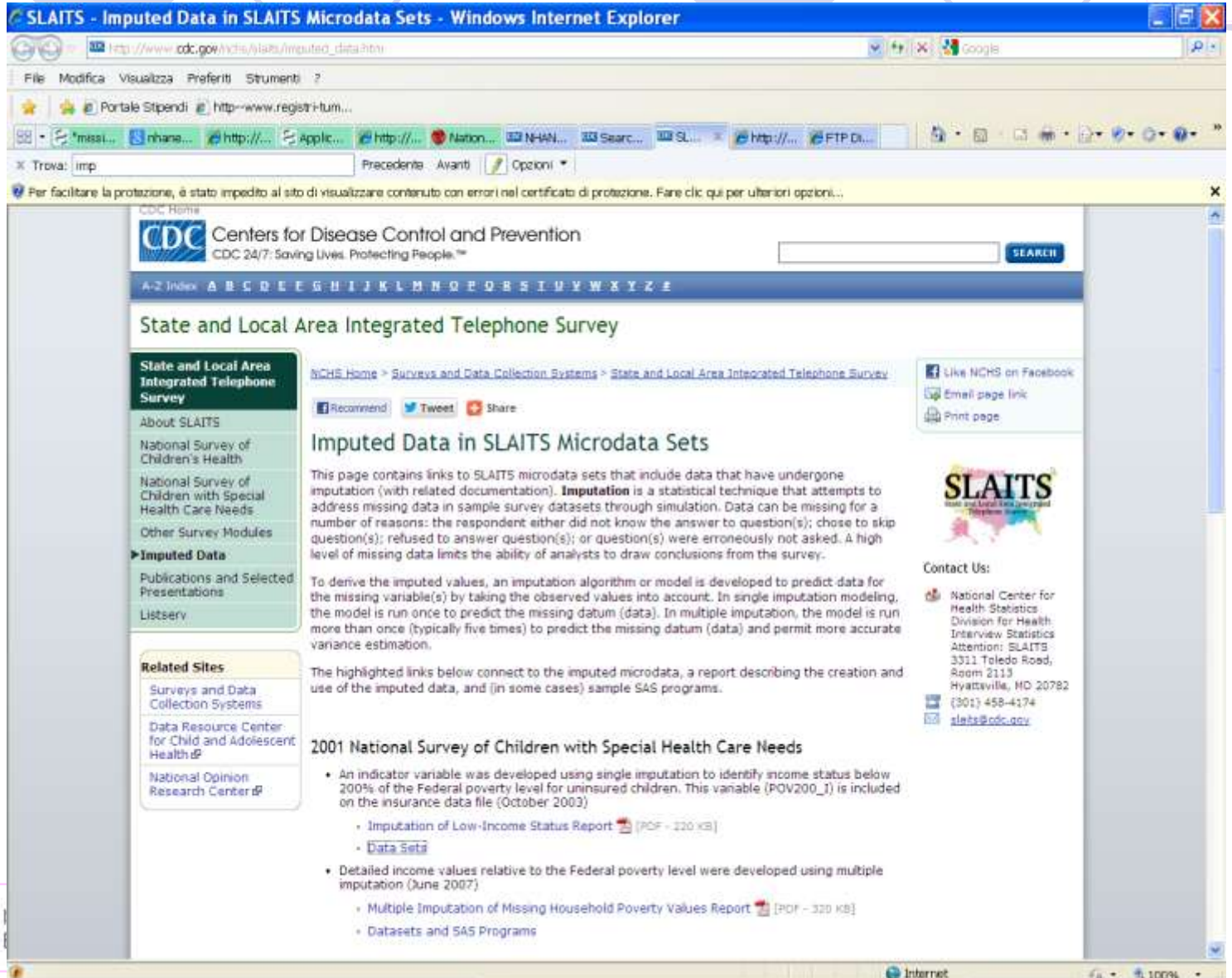
expectation; imputation; missing data; probability weighting

"Theory suggests that, if correctly and thoughtfully applied, **imputation methods should reduce bias and increase precision** in everyday use. We were **unable to assess the impact** of these methods in practice because of the **rarity of use and lack of detail in description**".



"We remain hopeful that inclusion of our suggested details in future publications will demonstrate to the field at large how imputation can reduce bias and increase precision in everyday use and that investigators will become more likely to utilize these methods".

http://www.cdc.gov/nchs/slaits/imputed_data.htm

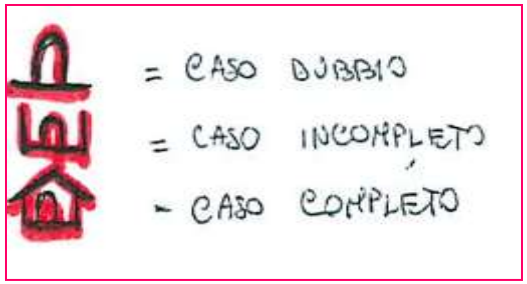


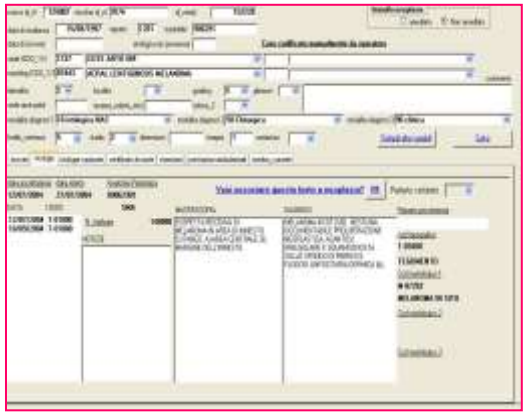
- Alto Adige CR (Guido Mazzoleni)
- Ferrara CR (Stefano Ferretti)
- Friuli Venezia Giulia CR (Diego Serraino)
- Modena CR (Massimo Federico)
- Parma CR (Maria Michiara)
- Romagna CR (Fabio Falcini)
- Veneto CR (Paola Zambon)
- Biella CR (Adriano Giacomin)
- Como CR (Gemma Gola)
- Genova CR (Marina Vercelli)
- Milano CR (Mariangela Autelitano)
- Sondrio CR (Roberto Tessandori)
- Firenze-Prato CR (Emanuele Crocetti)
- Latina CR (Fabio Pannozzo)
- Napoli CR (Mariao Fusco)
- Palermo CR (Adele Traina, Giuseppe Carruba)
- Ragusa CR (Rosario Tumino)
- Siracusa CR (Maria Lia Contrino, Anselmo Madeddu)

- Airtum Working Group
- Università degli Studi di Torino
- Fondazione ISI, Torino
- Emanuele Crocetti
- Enzo Coviello
- Costanza Pizzi

Background: Italian Cancer Registries

- ✓Incidence and risk factor — 1980
- ✓Prevalence and survival — 1990
- ✓ impact of screening programs — 2000
- ✓Diagnostic and therapeutic iter's evaluation — 2010





- ✓ Hospital discharge records
- ✓ Pathology records
- ✓ Death certificates

…

- ✓ Oncology, radiology and
- ✓ Haematology outpatients data
- ✓ Specialist services
- ✓ Exemptions from payment
- ✓ Pharmacology databases
- ✓ Screening services' information
- ✓ Reimbursement (treatment abroad)
- ✓ Palliative care services' data
- ✓ Diagnostic imaging services

ispo E LA PREVENZIONE ONCOLOGICA

Metodo delle imputazioni multiple

- ***Imputations***
  - Generate m incomplete copies of the data-sets, replacing missing values with values extracted from the posterior distribution of the data

- ***Analisys***
  - Of each data-set separately

- ***Pooling***
  - *Point estimates*
  - *Standard errors*.

Comandi utili in STATA e note sulla analisi di sopravvienza

Net survival
(to breast cancer)

Cause specific survival
event = death due to breast cancer

Relative survival
event = death
- ✓qualità definizione causa
- ✓cause multiple

All causes of death
Study population → S_o

Reference population → S_e

$$\frac{S_o}{S_e}$$

Instantaneous death rate in time t, measured among alive patients

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t \mid T \geq t)}{\Delta t} = \frac{d \ln[S(t)]}{dt}$$

[0;+∞)

+ flexibility

excess mortality

$$\nu(t; z) - \exp(\alpha + X\beta)z)$$

Dickman PW et al. Regression models for relative survival. Stat. Med., 2004 (23); 51–64.

Application: Results

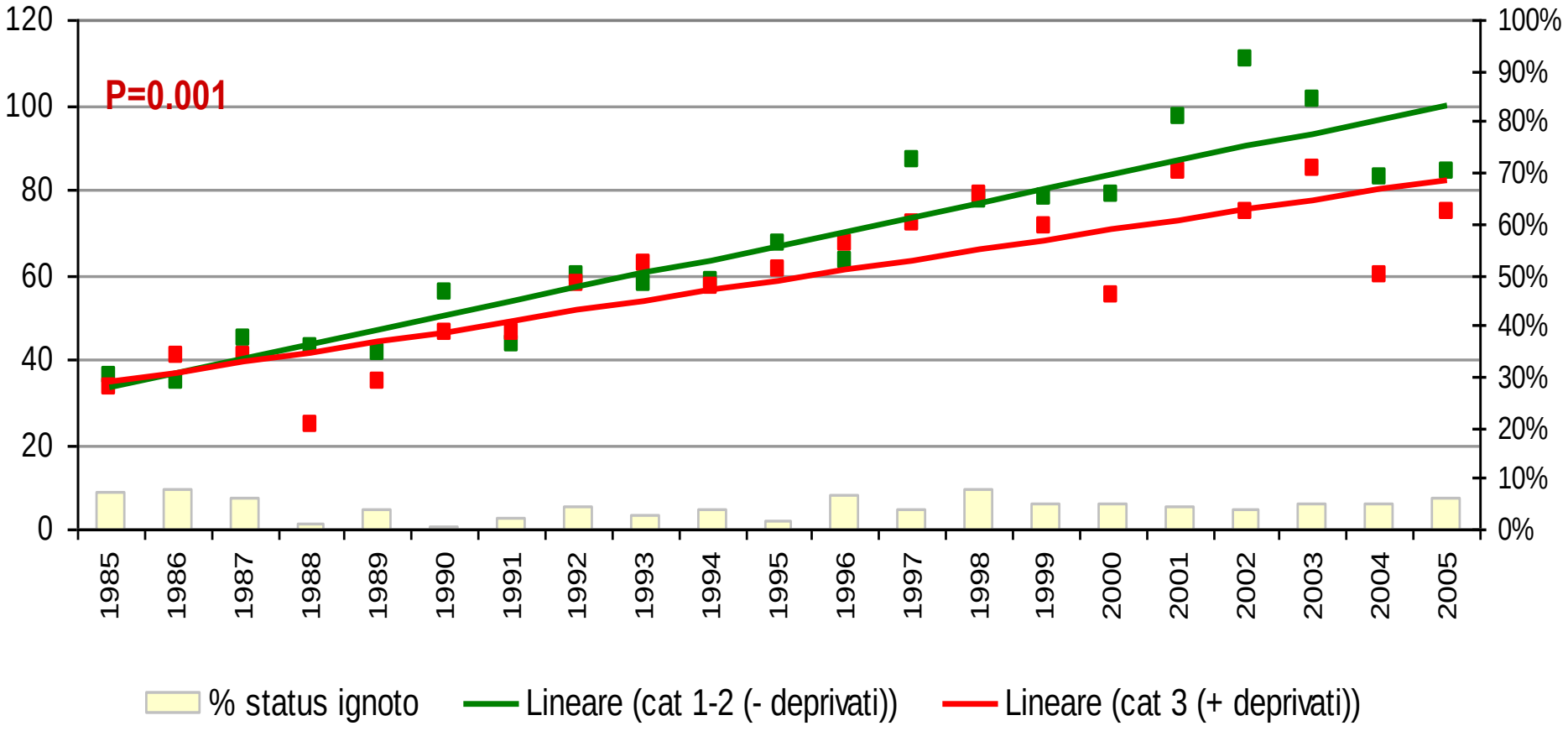
		Cases with complete information in stage variable					Complete analysis (all cases)					Cases, after imputation procedure				
Area	stage	EHR	Err. Std.	95% CI		p-value	EHR	Err. Std.	95% CI		p-value	EHR	Err. Std.	95% CI		p-value
AIRTUM-p	I (ref.)	1					1					1				
	IIA	5.02	0.54	4.07	6.18		5.48	0.66	4.33	6.94		5.08	0.51	4.17	6.2	
	IIB	12.79	1.28	10.5	15.57		13.98	1.6	11.17	17.49		13	1.28	10.7	15.78	
	IIIA	20.28	2.34	16.18	25.41		22.21	2.86	17.26	28.6		20.96	2.1	17.21	25.53	
	IIIB	24.91	3.29	19.23	32.26		25.99	3.55	19.88	33.98		26.63	2.9	21.5	32.98	
	IIIC	34.44	4.09	27.29	43.46		36.94	4.97	28.38	48.1		37.29	4.76	29.04	47.9	
	IV	118.18	13.28	94.81	147.3		112.45	13.83	88.36	143.12		78.49	8.76	63.05	97.73	
						< 0.0000	18.81	2.47	14.54	24.34	< 0.0000					< 0.0000

AIRTUM-pool and geographic areas. Five-years relative survival, Excess-hazard ratio for stage (reference stage = 1), adjusted by follow-up time, period of diagnosis, age group, cancer registry on *(a)* the cases with complete information for stage variable, *(b)* all cases, considering missing in stage variable as a valid category - missing category method, *(c)* on all cases, after applying multiple imputation

ISTITUTO PER LO STUDIO E LA PREVENZIONE ONCOLOGICA

PROSTATA

Comune di Firenze. Tumore della prostata. Tassi di incidenza std pop. EU per status socioeconomico



Definitions (Rubin)

"It is not possible to distinguish between MAR and MNAR only on the basis of the observed data; however collecting and including in the analyses the variables potentially associated with the outcomes, makes the assumption MAR more plausible"

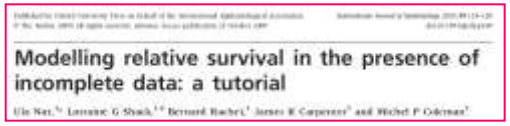
Rubin, DB, Multiple Imputation for Nonresponse in Surveys, 1987, New York, Wiley

Metodi ad-hoc per il trattamento dei dati mancanti

- Complete records analysis
- Missing category method
- Mean imputation
- Regression (conditional) mean imputation
- Random or stochastic regression imputation
- Multiple imputation (MI)
- Inverse probability weighting

Metodi ad-hoc per il trattamento dei dati mancanti

Table 2 Adjusted[a] EHR of death within 1 and 5 years of diagnosis, after use of different methods to impute missing data values: adults diagnosed with colorectal cancer during 1997–2004 in the North West Region of England


Modelling relative survival in the presence of incomplete data: a tutorial

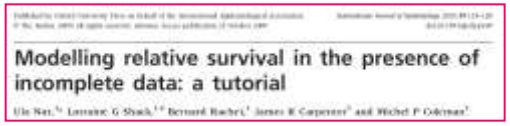
complete case analysis

	EHR	95% CI	
Stage			
I	**1.00**	–	–
II	**3.56**	2.69	4.72
III	**10.20**	7.72	13.48
IV	**26.39**	19.60	35.53
Missing			
Morphology			
Adenocarcinoma	**1.00**	–	–
Mucinous and serous	**1.03**	0.94	1.13
Other	**1.17**	0.61	2.24
Neoplasm NOS			
Grade			
I	**1.00**	–	–
II	**1.18**	1.07	1.30
III/IV	**2.04**	1.82	2.28
Missing			
Colorectal site (ICD–10 code)			
Colon (C18)	**1.00**	–	–
Rectosigmoid (C19)	**0.82**	0.73	0.91
Rectum (C20)	**0.76**	0.70	0.82
Sex			
Male	**1.00**	–	–
Female	**0.93**	0.88	0.99

Variable	Patients	
	N	%
	29 563	**100**
Stage		
I	2193	7.4
II	7326	24.8
III	7726	26.1
IV	643	2.2
Missing	11 684	39.5

Metodi ad-hoc per il trattamento dei dati mancanti

Table 2 Adjusted[a] EHR of death within 1 and 5 years of diagnosis, after use of different methods to impute missing data values: adults diagnosed with colorectal cancer during 1997–2004 in the North West Region of England


Modelling relative survival in the presence of incomplete data: a tutorial

missing category method

	EHR	95% CI	
Stage			
I	**1.00**	–	–
II	**3.10**	2.39	4.03
III	**8.35**	6.46	10.80
IV	**19.12**	14.56	25.11
Missing	**11.54**	8.94	14.91
Morphology			
Adenocarcinoma	**1.00**	–	–
Mucinous and serous	**1.12**	1.04	1.21
Other	**1.29**	0.98	1.68
Neoplasm NOS	**2.71**	2.52	2.90
Grade			
I	**1.00**	–	–
II	**1.26**	1.16	1.37
III/IV	**2.31**	2.10	2.55
Missing	**1.50**	1.37	1.65
Colorectal site (ICD–10			
Colon (C18)	**1.00**	–	–
Rectosigmoid (C19)	**0.93**	0.86	1.00
Rectum (C20)	**0.84**	0.81	0.88
Sex			
Male	**1.00**	–	–
Female	**0.96**	0.92	1.00

Variable	Patients	
	N	%
	29563	**100**
Stage		
I	2193	7.4
II	7326	24.8
III	7726	26.1
IV	643	2.2
Missing	11684	39.5