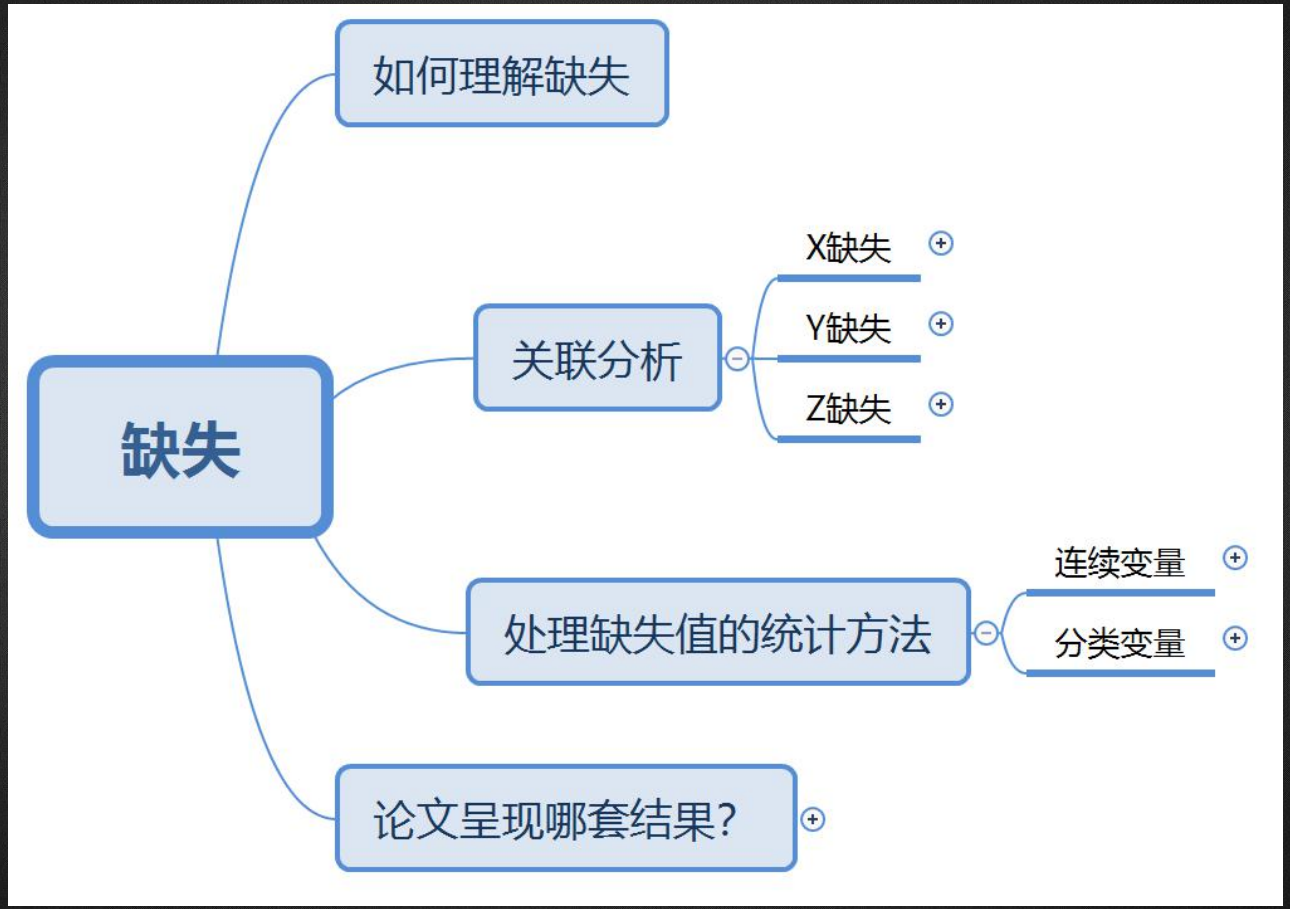
数据分析缺失值的常见处理方法

Empower U, Department of Epidemiology and Biostatistics
X&Y solutions Inc. in Boston

陈星霖 Ph.D/易佾学院

2021-10-9

提纲



如何理解缺失

有缺失才正常。

一项临床研究，尤其是样本量较大且涉及随访的研究，如果所有变量都不缺失，审稿人可能会怀疑数据造假。

常见误区与解决方案

常见误区

有数据缺失的研究对象都被研究者删除了。被质疑数据造假。

解决方案

需要学习如何采用合理的方法处理缺失数据。
纳排标准制定、统计分析（处理缺失值）、研究设计（防止偏性）三个环节。

- 暴露因素（X）缺失，通常排除。
- 结局指标（Y）缺失，通常排除。有的研究不排除，分析偏性对结果的影响，见后文实例。
- 混杂因素（Z）缺失，不要剔除。需要描述缺失情况并且处理缺失值。回归方程调整混杂时，要求在每个变量都不缺失的研究对象中运行。

处理缺失值的统计方法

1. 增加哑变量：分类变量

Table 1 Characteristics of cases and controls

	Cases	Controls
N	44	220
Age (years) (±SD)	82.2 (6.7)	82.2 (6.6)
Smoking (%)		
Non-current smoker	77.3	70.9
Current smoker	2.3	3.2
Not recorded	20.5	25.9
Alcoholism	0.0	0.0
Body mass index (kg/m^2) (±SD)	29.4 (4.9)	29.1 (5.3)
<20 kg/m^2 (%)	0.0	1.4
20–<25 kg/m^2 (%)	9.1	14.1
25–<30 kg/m^2 (%)	29.6	25.0
>=30 kg/m^2 (%)	31.8	32.3
Not recorded (%)	29.6	27.1
Comorbidities (%)		
Previous fracture	20.5	8.2
Kidney disease	4.6	5.0
Malabsorption	2.3	1.4

研究服药和骨折的关系，BMI是混杂因素需要调整，从表1看出BMI缺失占1/3左右（not recorded：病例组29.6%，对照组27.1%）。

处理缺失的方法：把缺失BMI的做为一组进入回归方程。即分成五组：<20、≥20且<25、≥25且<30、≥30、缺失。

Erviti, J., et al., Oral bisphosphonates are associated with increased risk of subtrochanteric and diaphyseal fractures in elderly women: a nested case-control study. BMJ Open, 2013.

处理缺失值的统计方法

2. 增加哑变量：连续变量

	A	B	C	D
1	id	BMI	G	BMINEW
2	1	15	0	15
3	2	16	0	16
4	3		1	0
5	4	18	0	18
6	5		1	0

例如体质指数（BMI）缺失，可创建两个指标：

（1）二分类变量G：G=1表示BMI缺失，G=0表示BMI不缺失
（2）原始连续变量BMI中缺失的用0填补：BMInew

基本原理 回归分析时这两个指标需要同时在方程里： Y=b1*BMInew+b2*G

- 当BMI缺失时，BMInew=0、G=1，因此Y=b2
- 当BMI不缺失时，BMInew=BMI原始测量值、G=0，因此Y=b1*BMI

BMI缺失的研究对象对Y的作用体现在b2上；BMI不缺失的研究对象对Y的作用体现在b1上。

处理缺失值的统计方法

2. 增加哑变量：连续变量

优势

(1) 不损失样本量；

(2) 考虑到了BMI缺失对结果的影响，因为BMI缺失的研究对象对Y的作用体现在b2上；

(3) 数据分析用的是原始的真实数据，没有用插补等模拟的“假”数据。

统计分析撰写参考

Dummy variables were used to indicate missing covariate values.

Dummy variables were used to indicate missing covariate values. For missing information on pack-years of smoking, the median among smokers was imputed; in case of missing BMI, information was carried forward once. On average, 9.5% of covariate information was missing across 24 years of follow-up.

Association Between Rotating Night Shift Work and Risk of Coronary Heart Disease Among Women. JAMA, 2016.

如何理解缺失处理

精细

连续变量与分类变量缺失：考虑到其他指标的关系后填补

粗略

连续变量缺失：用中位数、均数填补

分类变量缺失：用一个新分类填补

处理缺失值的统计方法

3. 多重插补（MI，Multiple imputation）

例1. 统计学方法部分：教育、BMI和体力活动的变量有缺失，我们采用多重插补（基于SAS的链式方程）生成了5套数据，对数据缺失进行处理。

clusively decaffeinated coffee. We used multiple imputation, based on 5 replications and the Markov-chain Monte Carlo method in the SAS MI procedure, to account for missing data on education, BMI, and physical activity. This resulted in the exclusion of 30 participants for whom imputation was not possible. We also performed sensitivity analyses using a complete-case analysis (n = 174 579).

2017年发表在Ann Intern Med（SCI影响因子17.1分）的喝咖啡与低死亡风险的研究。Association of Coffee Consumption With Total and Cause-Specific Mortality Among Nonwhite Populations.

处理缺失值的统计方法

3. 多重插补（MI，Multiple imputation）

例2. 统计学方法部分：为避免直接排除缺失值带来的统计检验效能减少和偏性，使用多重插补来估算缺失值。

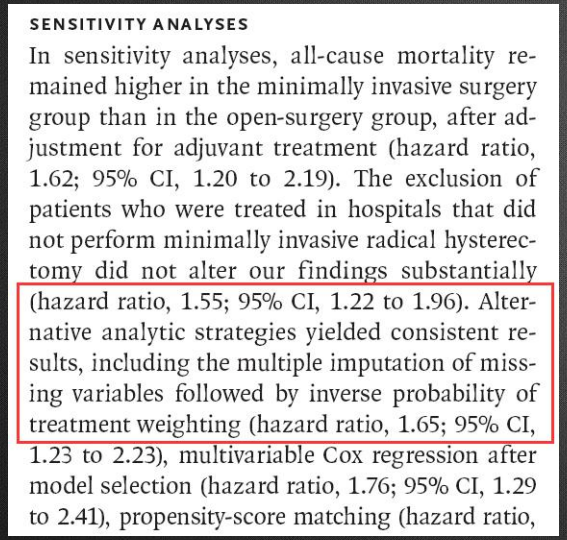
To maximise statistical power and minimise bias that might occur if women with missing data were excluded from analyses, we used multivariate multiple imputation with chained equations to impute missing values[24] (appendix). We repeated all analyses with the complete data cohort for comparison. The appendix gives additional details of the statistical analyses.

2015年发表在Lancet Diabetes Endocrinol（SCI影响因子25.3分）的高血糖与不良围产期结局的研究。Association between hyperglycaemia and adverse perinatal outcomes in south Asian and white British women: analysis of data from the Born in Bradford cohort.

处理缺失值的统计方法

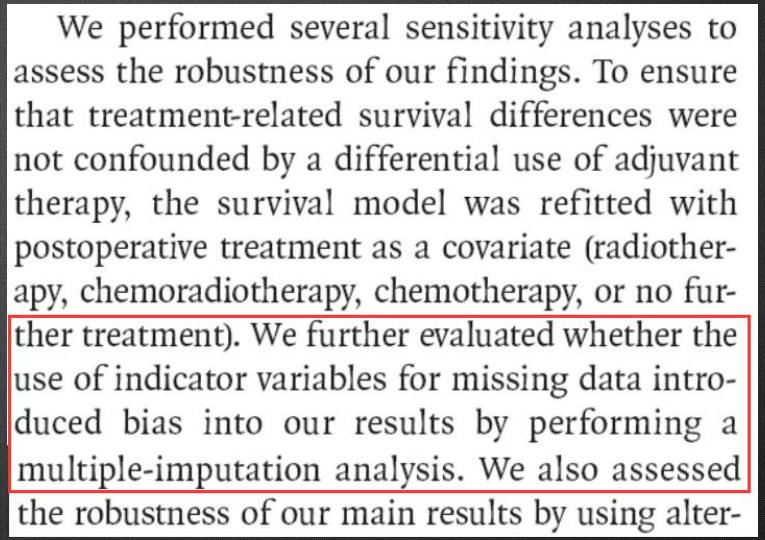
3. 多重插补（MI，Multiple imputation）

例3. 缺失数据的多重插补写在统计分析的敏感性分析部分。并在结果中呼应：考虑到缺失数据的处理后，结果一致。



SENSITIVITY ANALYSES

In sensitivity analyses, all-cause mortality remained higher in the minimally invasive surgery group than in the open-surgery group, after adjustment for adjuvant treatment (hazard ratio, 1.62; 95% CI, 1.20 to 2.19). The exclusion of patients who were treated in hospitals that did not perform minimally invasive radical hysterectomy did not alter our findings substantially (hazard ratio, 1.55; 95% CI, 1.22 to 1.96). Alternative analytic strategies yielded consistent results, including the multiple imputation of missing variables followed by inverse probability of treatment weighting (hazard ratio, 1.65; 95% CI, 1.23 to 2.23), multivariable Cox regression after model selection (hazard ratio, 1.76; 95% CI, 1.29 to 2.41), propensity-score matching (hazard ratio,



We performed several sensitivity analyses to assess the robustness of our findings. To ensure that treatment-related survival differences were not confounded by a differential use of adjuvant therapy, the survival model was refitted with postoperative treatment as a covariate (radiotherapy, chemoradiotherapy, chemotherapy, or no further treatment). We further evaluated whether the use of indicator variables for missing data introduced bias into our results by performing a multiple-imputation analysis. We also assessed the robustness of our main results by using alter-

2018年发表在NEJM的早期宫颈癌微创根治性子宫切除术后的存活率的研究. Melamed, A., et al., Survival after Minimally Invasive Radical Hysterectomy for Early-Stage Cervical Cancer. New England Journal of Medicine, 2018.

正文中放哪套结果?

原始数据是真的，插补后的是假的。

如果真的和假的核心结果一致，正文放真的，附表放假的。或者同时呈现在正文。

注意一定要在正文呈现原始数据，否则可能被审稿人认为作者挑选数据。

支持证据放附表

Supplementary tables and figures

Table S1: Distributions of variables with missing data comparing observed complete case data to results from pooling the datasets with imputed variables from multiple imputation

	Level/Unit	Number (%) with missing data	Complete case	Multiple imputation
Birthweight	Mean (se) standard deviation score	1	-0·37 (0·01)	-0·37 (0·01)
Sum of skinfolds	Mean (se)	3051 (32·1)	9·82 (0·03)	9·75 (0·02)
Pre-eclampsia	%	389 (4·1)	2·5	2·5
Instrumental vaginal delivery[a]	%	7 (0·1)	12·4	12·4
Maternal BMI	Mean (se)	436 (4·6)	25·8 (0·06)	25·9 (0·06)
Maternal education	% 5+ GCSE equivalent	126 (1·3)	31·5	31·5
	% Higher than A-level equivalent		25·6	25·6
Smoking	%	15 (0·2)	17·0	17·0
Alcohol	%	36 (0·4)	20·6	20·6
Parity	% primiparious	358 (3·8)	41·7	41·4
Family history of diabetes	%	297 (3·1)	25·1	25·1
Family history of hypertension	%	306 (3·2)	27·4	27·4
Previous macrosomia	%	874 (16·4)	4·5	4·8

[a]These analyses exclude women who had a Caesarean section, therefore the N=7526.

通过对比可以看出原始数据和插补后的数据变量分布情况基本一致。

这个表不是文章的主要结果，所以没有放在正文中，放在附表中做为支持证据呈现。

Farrar, D., et al., Association between hyperglycaemia and adverse perinatal outcomes in south Asian and white British women: analysis of data from the Born in Bradford cohort. Lancet Diabetes Endocrinol, 2015. 3(10): p. 795-804.

深入分析：偏性

需要考虑Z缺失是否影响核心结果，使结果偏向阳性还是阴性。
如研究服药和骨折的关系，BMI缺失。假如缺失BMI的是偏向于服药但没有骨折的人，除掉了这些人会使结果偏向于：服药的人发生骨折发生率高。

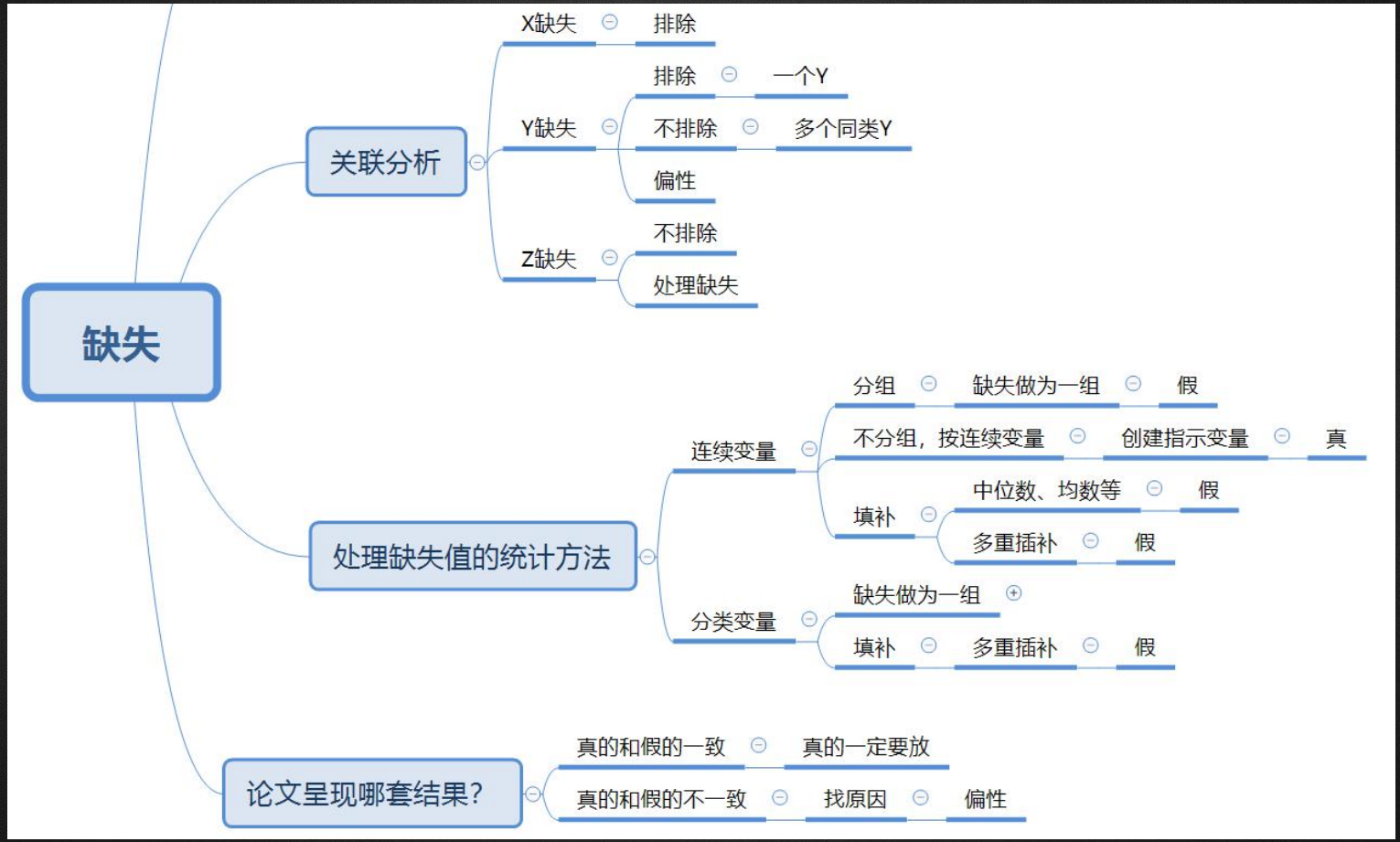
解决方案：按BMI是否缺失分成两个组，看其他变量的分布，发现BMI缺失的73个人年龄偏大，服药的比例小（8.2%）并且骨折的比例稍大（17.8%）。因此判断缺失的人中服药带来的骨折风险较小。

- 如果把缺失的人加上会导致结果偏向阴性：即服药的人不容易发生骨折。
- 然而目前文章结果是阳性的，表明如果BMI缺失的人排除，本研究的阳性发现作用更强。因此BMI缺失的问题，不会改变本研究核心结果。

BMI categorical recoded	0	1
N	191	73
AGE	81.2 ± 6.3	84.7 ± 6.8
Y		
0	160 (83.8%)	60 (82.2%)
1	31 (16.2%)	13 (17.8%)
X		
0	161 (84.3%)	67 (91.8%)
1	30 (15.7%)	6 (8.2%)

Erviti, J., et al., Oral bisphosphonates are associated with increased risk of subtrochanteric and diaphyseal fractures in elderly women: a nested case-control study. BMJ Open, 2013.

小结


缺失
关联分析
X缺失
排除
Y缺失
排除
一个Y
不排除
多个同类Y
偏性
Z缺失
不排除
处理缺失
处理缺失值的统计方法
连续变量
分组
缺失做为一组
假
不分组，按连续变量
创建指示变量
真
填补
中位数、均数等
假
多重插补
假
分类变量
缺失做为一组
填补
多重插补
假
论文呈现哪套结果?
真的和假的一致
真的一定要放
真的和假的不一致
找原因
偏性

Thank you for your time!