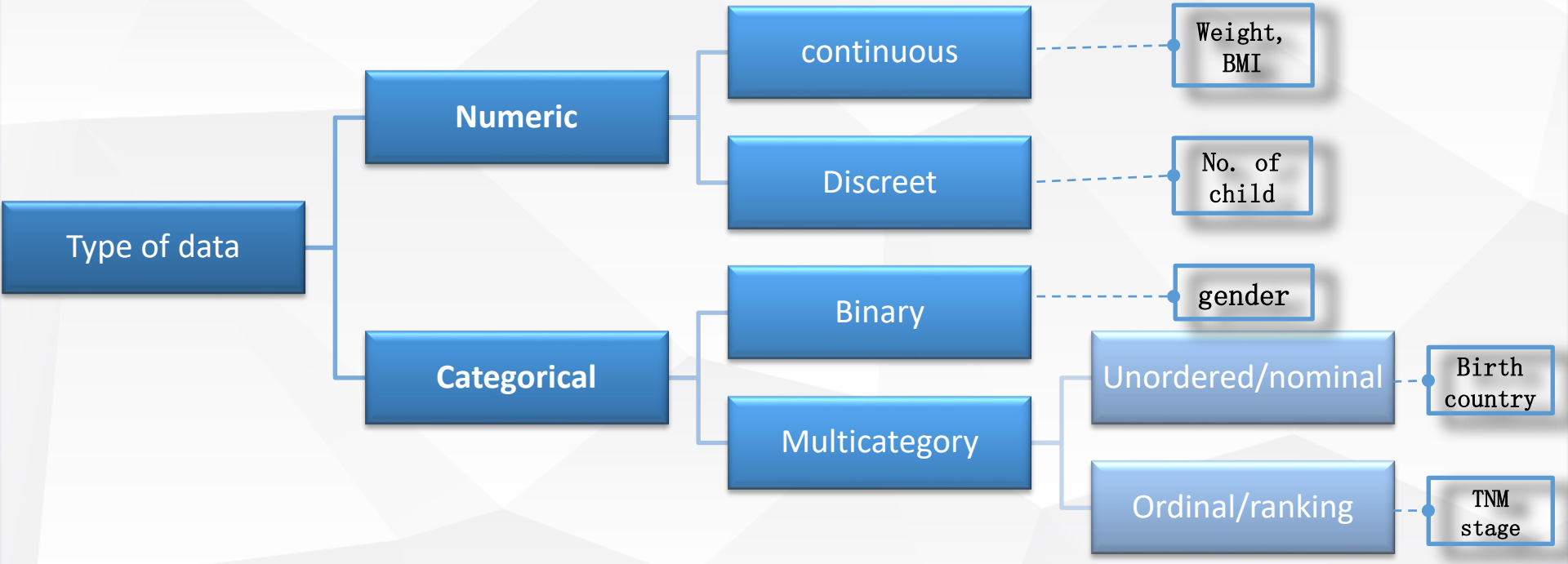
Selection of Appropriate Statistical Methods for data analysis

- Basic conception
- Choose method
- Sample size
- Questions?

Basic conception

Basic Concepts

1.1 type of data



1.1 type of data-dummy variable

Ordinal

Score	Fail（<60）	Pass（60-75）	Good（75-85）	Excellent（>85）
Code	1	2	3	4

nominal

Profession	Student	Farmer	Worker	Civil servant	Other
Code	1	2	3	4	5

In regression model, nominal variable should be transformed into dummy variable：

Set “other” as a reference，create four dummy variables (X1-X4)

	x1	x2	x3	x4
Student	1	0	0	0
Farmer	0	1	0	0
Worker	0	0	1	0
Civil servant	0	0	0	1
Other	0	0	0	0

1.1 type of data-survival data

time-to-event data： combination of status (categorical) and time (continuous)

Follow-up of lung cancer patients

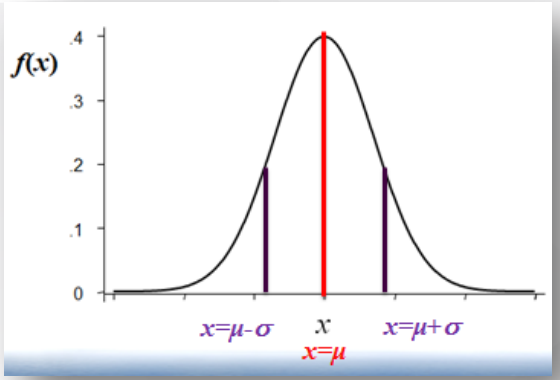
Patient ID	(Survival) Status	(Survival) Days
1	Alive	1455
2	Death	90
3	Alive	1390
4	Death	1129
5	Death	983

Neither x2 test nor t-test can use all the information

1.2 distribution of data

Parametric test

Height

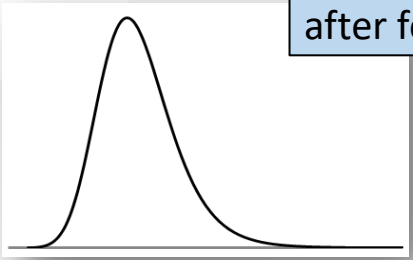


a. Normal distribution

$$X \sim N(\mu, \sigma^2)$$

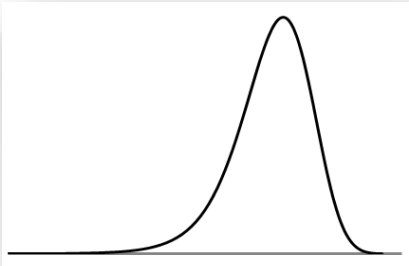
Non-parametric test

Symptom occurrence after food poisoning



b. Right skewed data

Diagnostic age of tumor



c. Left skewed data

1.3 Study design

◆Factor： are the variables (treatments) in the study that we believe will influence the results

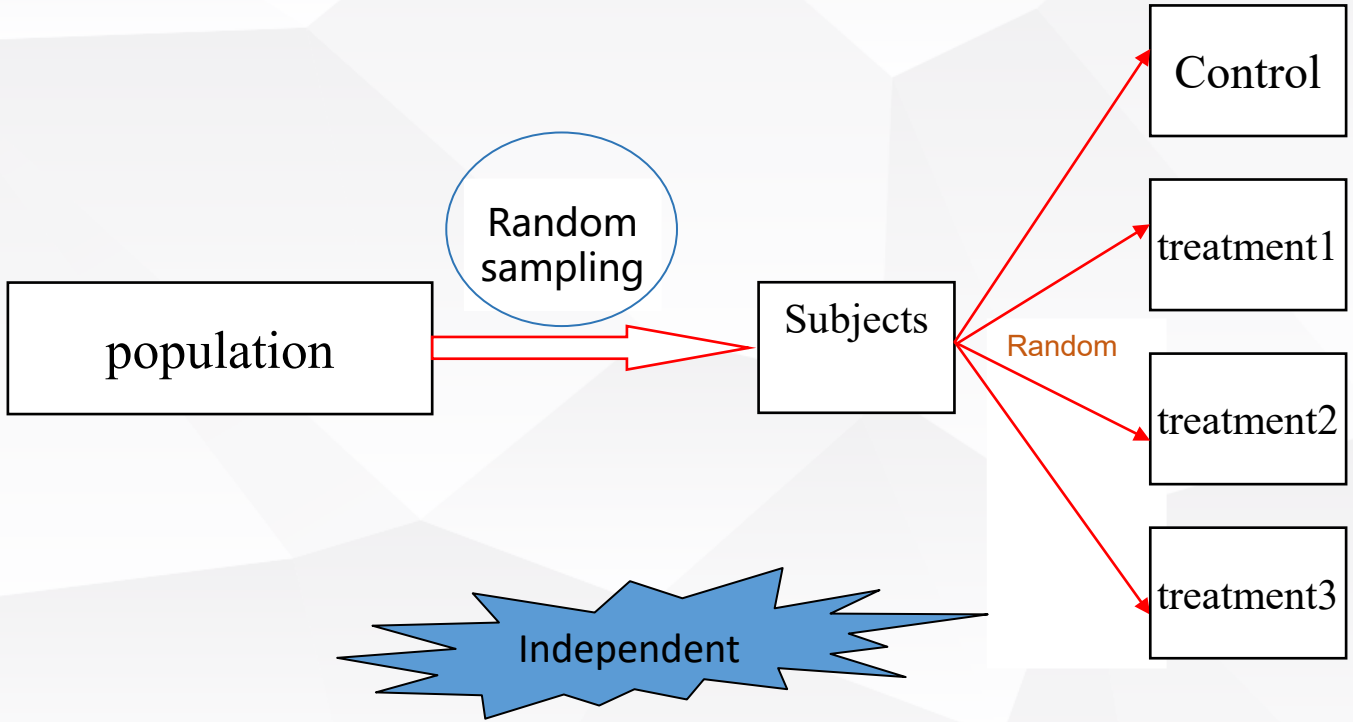
◆level： are the “values” of that factor in an experiment

Example: antihypertensive drug

Group	
Experimental	Drug: 1 unit/d
	Drug: 3 unit/d
Control	Placebo

1.3 Study design-completely random design

（1）Completely random design

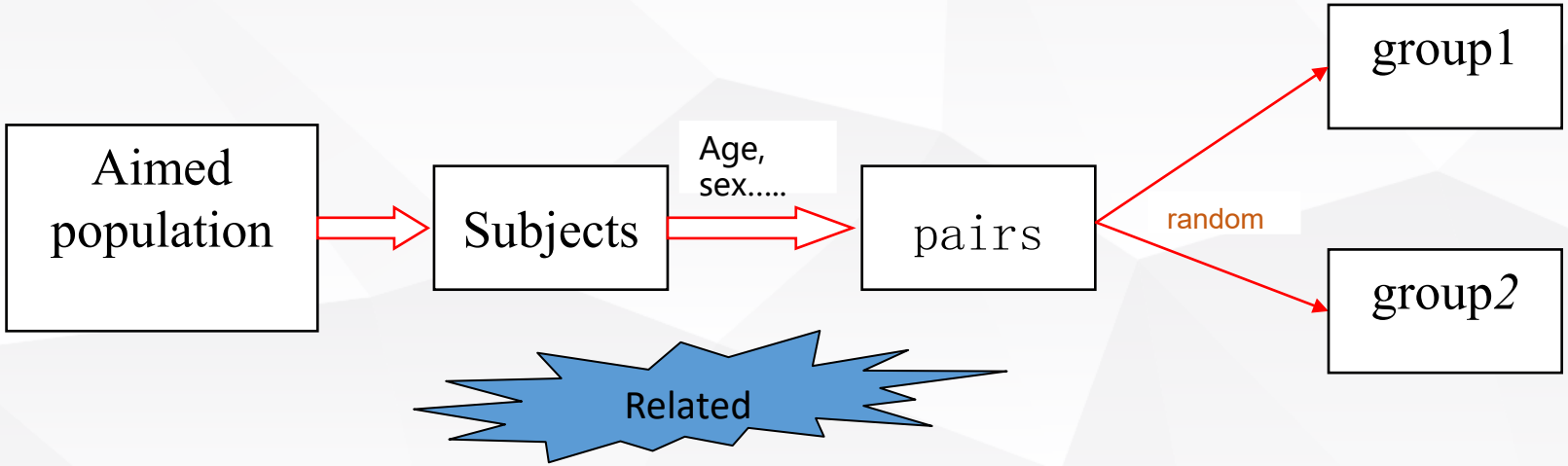


1.3 Study design-paired study

（2）Paired study

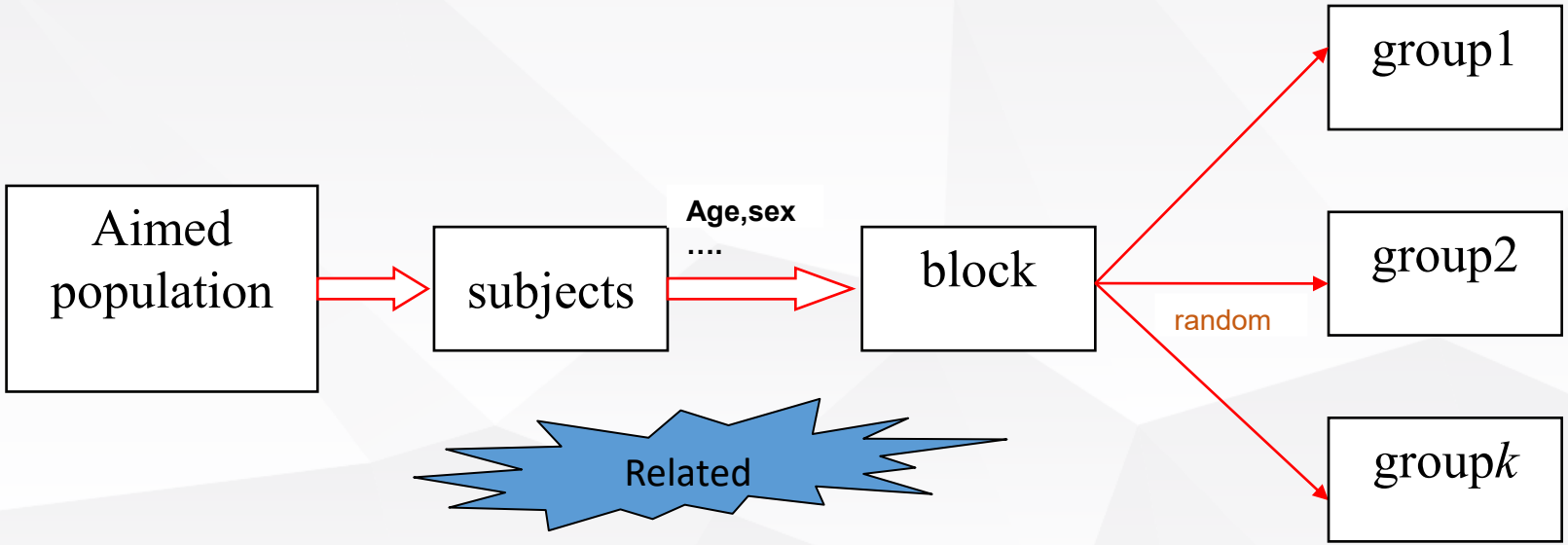
①Measuring the same object at two different time points (Duplicate, Pre-Post Measurements);
②Different parts of the same object;
③Measured the same object with two different methods.

④Matching.



1.3 Study design-RBD

（3）randomized block design

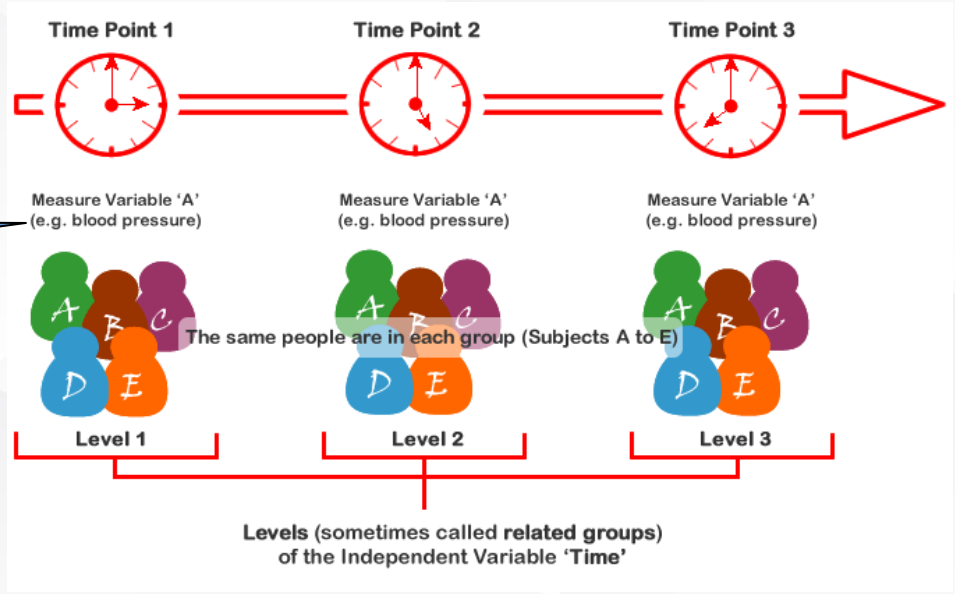


1.3 Study design-RMD

（4）Repeated measurement design

multiple measurements of each subject, extension of paired study

Related

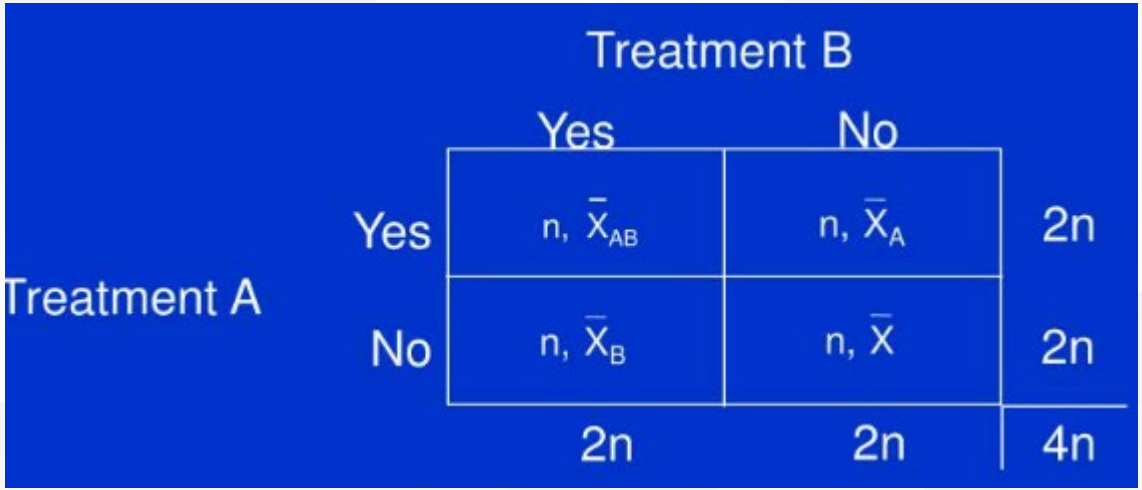


1.3 Study design-factorial design

（5）Factorial design

multiple measurements of each subject, extension of paired study

2×2 factorial design



Choose from common methods

These are general guidelines and should not be construed as hard and fast rules

2.1 Factors Influencing Selection of Statistical Methods

1 Aim and objective of the study	2 Type and distribution of the data used	3 Study design
Descriptive Inferential: Compare? Relationship? Predict?	Continuous Normal? Categorical ordinal/nominal? Survival	Independent/related? Sample size? Factors/levels?

2.2 Descriptive

①aim → ②data → ③design

Aim	Data			
	Continuous Normal data	Ordinal, continuous skewed data	Nominal	Survival data
Descriptive	Mean±SD	Median (IQR)	N (%)*	Life table, Kaplan-Meier survival curve

* proportion, rate, percent….

Table 1. Baseline Characteristics of the Trial Participants

Characteristics	Antihypertensive Treatment (n = 2038)	Control (n = 2033)
Age, mean (SD), y	62.1 (10.8)	61.8 (11.0)
Men, No. (%)	1317 (64.6)	1287 (63.3)
Time from onset to randomization, mean (SD), h	15.3 (12.9)	14.9 (13.0)
Blood pressure at entry, mean (SD), mm Hg		
Systolic	166.7 (17.3)	165.6 (16.5)
Diastolic	96.8 (10.8)	96.5 (11.4)
Body mass index, mean (SD)[a]	24.9 (3.2)	25.0 (3.1)
NIHSS score, median (IQR)[b]	4.0 (2.0-7.0)	4.0 (3.0-8.0)
History of hypertension, No. (%)	1610 (79.0)	1599 (78.7)
Current use of antihypertensive medications, No. (%)	1014 (49.8)	983 (48.4)
Hyperlipidemia, No. (%)	137 (6.7)	140 (6.9)
Diabetes mellitus, No. (%)	369 (18.1)	350 (17.2)
Coronary heart disease, No. (%)	216 (10.6)	228 (11.2)
Current cigarette smoking, No. (%)	725 (35.6)	760 (37.4)
Current alcohol drinking, No. (%)	614 (30.1)	639 (31.4)
Ischemic stroke subtype, No. (%)[c]		
Thrombotic	1575 (77.3)	1595 (78.5)
Embolic	99 (4.9)	103 (5.1)
Lacunar	417 (20.5)	385 (18.9)

Abbreviations: IQR, interquartile range; NIHSS, National Institutes of Health Stroke Scale.

age， continuous normal data

NIHSS scale, ranked data

DM, categorical data

PMID: 24240777

2.3 inferential

①aim → ②data → ③design

Aim	Data			
	Continuous Normal data	**Ordinal, continuous skewed data**	**Nominal**	**Survival data**
Compared with hypothetical value	One sample t-test	One sample Wilcoxon signed rank test	Chi-square goodness-of-fit test	Kaplan-Meier survival curve

2.3 inferential

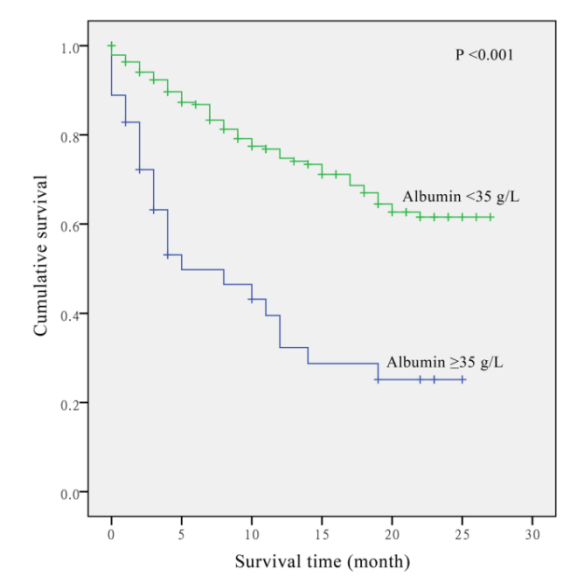
①aim → ②data → ③design

Aim, Design	Data			
	Continuous Normal data	**Ordinal, continuous skewed data**	**Nominal**	**Survival data**
Compare between Two unpaired groups	Independent samples t-test	Mann Whitney U test/Wilcoxon rank sum test	Chi-square test¥	Log-rank test
Compare between Two paired groups (paired design)	Paired t-test	Related samples Wilcoxon signed-rank test	McNemar's test*	Conditional proportional hazards regression

¥ n<40 or E<1, use Fisher's exact test

*McNemar's test is only suitable for 2✕2 contingency table

Variables	Total (N = 387)	Albumin <35 g/L (N = 54)	Albumin ≥35 g/L (N = 333)	*P*-value
Age (years)	387	63.83 ± 10.94	62.16 ± 10.37	0.276
Sex				0.002
Male sex	241	44 (81.5)	197 (59.2)	
Female sex	146	10 (18.5)	136 (40.8)	
Smoking				0.005
Never smoker	183	15 (27.8)	168 (50.5)	
Former smoker	102	17 (31.5)	85 (25.5)	
Current smoker	102	22 (40.7)	80 (24.0)	
BMI (kg/m^2)	387	21.70 ± 2.83	22.81 ± 3.41	0.024
Tumor type				0.852
AC	259	35 (64.8)	224 (67.3)	
SCC				
unknown				
Cancer stage				0.170



Post-diagnostic C-reactive protein and albumin predict survival in Chinese patients with non-small cell lung cancer: a prospective cohort study.

Yang JR, Xu JY, Chen GC, Yu N, Yang J, Zeng DX, Gu MJ, Li DP, Zhang YS, Qin LQ.

Continuous variables were expressed as the means with the standard deviation, and were compared using the Student's t test. The Chi-Square or Mann-Whitney U test was used to compare the categorical variables, which was presented as the number and percentage of patients.

Survival analysis was performed using the Kaplan-Meier method, and the differences were assessed using the Log-Rank test.

2.3 inferential

①aim → ②data → ③design

Aim, Design	Data			
	Continuous Normal data	**Ordinal, continuous skewed data**	**Nominal**	**Survival data**
≥3 unpaired groups（i.e., CRD）	One way ANOVA$	Kruskal-Wallis test	Chi-square test	Log-rank test cox proportional hazards regression
≥3 Related groups（i.e., RBD）	ANOVA for RBD*	Friedman test	Cochrane Q test	Conditional Cox proportional hazards regression

If reject H0, post hoc analysis

*Repeated measures ANOVA

$ ANOVA for factorial design

Repeated measures logistic regression仅限二分类变量。

Increased Growth Differentiation Factor 15 Is Associated with Unfavorable Clinical Outcomes of Acute Ischemic Stroke.

Table 1. Characteristics of participants according to categories of serum GDF-15.[a]

Characteristics	Total	GDF-15, ng/L			*P* value
		<1200	1200-1800	>1800	
Number of subjects, n (%)	3066	1979 (64.54)	635 (20.71)	452 (14.74)	
Demographics					
Age, years	62.3 ± 10.8	59.3 ± 10.0	66.7 ± 9.6	69.7 ± 10.3	<0.001
Male, n (%)	1966 (64.12)	1270 (64.17)	405 (63.78)	291 (64.38)	0.994
Current cigarette smoking	1124 (36.66)	741 (37.44)	221 (34.80)	162 (35.84)	0.330
Current alcohol drinking	958 (31.25)	686 (34.66)	165 (25.98)	107 (23.67)	<0.001

Baseline characteristics were compared among these 3 groups using ANOVA or Chi-square test, when appropriate.

Kaplan-Meier survival curves, log-rank tests, and Cox proportional hazards models were used to evaluate the associations between these 3 groups and the cumulative incidence of death, cardiovascular events, and stroke recurrence.

2.3 inferential

①aim → ②data → ③design

Aim	Data			
	Continuous Normal data	Ordinal, continuous skewed data	Nominal	Survival data
Degree of linear relationship between two variables	Pearson correlation	Spearman correlation	Association analyses[&]	

[&] Phi coefficient and Kappa.

Scatter plot at first!

2.3 inferential

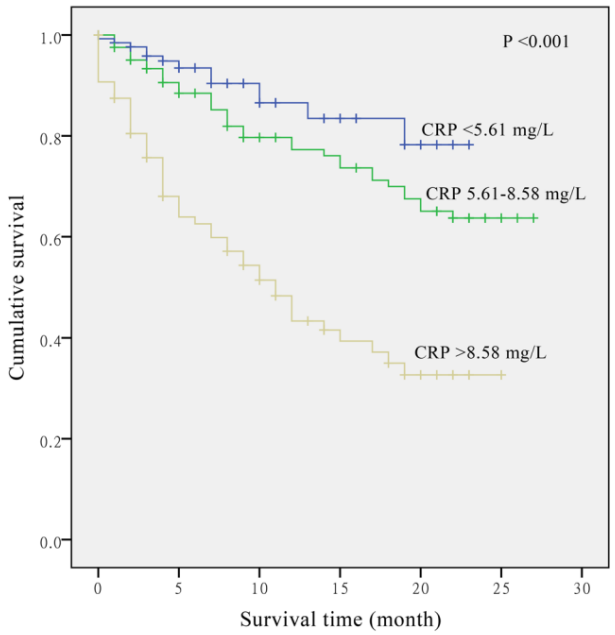
Aim	Data type of Outcome			
	Continuous Normal data	**Ordinal, continuous skewed data**	**Nominal**	**Survival data**
Predict 1 outcome variable by 1 independent variable	Simple linear regression	Nonparametric regression	(conditional) Logistic regression	Cox proportional hazards regression
Predict 1 outcome variable by ≥1 independent variable	Multiple linear regression; analysis of covariance		(conditional) Multiple logistic regression	Cox proportional hazards regression

Technically, assumptions of normality concern the errors rather than the dependent variable itself !

Many of these models produce estimates that are robust to violation of the assumption of normality, particularly in large samples.

Variables	N	Model 1		Model 2		Model 3	
		HR (95% CI)	*P*-trend	HR (95% CI)	*P*-trend	HR (95% CI)	*P*-trend
Multivariate analysis for CRP and albumin							
CRP (mg/L)			<0.001		0.001		0.003
<5.61	134	1		1		1	
5.61–8.58	124	1.61 (0.83–3.10)		1.59 (0.82–3.09)		1.56 (0.80–3.04)	
>8.58	129	3.32 (1.73–6.40)		2.84 (1.46–5.49)		2.64 (1.35–5.16)	
Albumin (g/L)			0.006		0.002		0.011
<35	54	1		1		1	
≥35	333	0.51 (0.31–0.82)		0.45 (0.27–0.74)		0.50 (0.29–0.85)	
Multivariate analysis for CRP/Alb ratio							
CRP/Alb ratio			<0.001		<0.001		<0.001
<0.14	148	1		1		1	
0.14–0.22	110	2.30 (1.18–4.49)		2.16 (1.10–4.28)		2.19 (1.11–4.34)	
>0.22	129	5.16 (2.73–9.78)		4.68 (2.45–8.95)		4.14 (2.15–7.98)	

Table 3. COX proportional hazards regression overall model of CRP, albumin and CRP/Alb ratio. Alb, albumin; HR, hazard ratio; CI, confidence interval; CRP, C-reactive protein; Model 1 includes age at baseline interview, sex, body mass index, family history of cancer, patient history of chronic obstructive pulmonary disease, smoking status and drinking habit. Model 2: model 1 plus tumor type, cancer stage and treatment. Model 3: model 2 plus history of chronic liver disease and white blood cell count.



- Hazard ratios (HR) and 95% confidence intervals (CI) of NSCLC death were estimated using the Cox proportional hazards regression model.

Post-diagnostic C-reactive protein and albumin predict survival in Chinese patients with non-small cell lung cancer: a prospective cohort study

2.3 inferential

If there are 2+ outcome variables, and the outcome variables are normal or interval data,
Then you can consider of one-way MANOVA, multivariate multiple linear regression, factor analysis and canonical correlation

Sample size

N

3.1 Small sample size

- For small sample size (average ≤15 observations per group), normality testing methods are less sensitive about non-normality and there is chance to detect normality despite having non-normal data. It is recommended that when sample size is small, only on highly normally distributed data, parametric method should be used otherwise corresponding nonparametric methods should be preferred.

3.2 sufficient sample size

- Similarly on sufficient or large sample size (average >15 observations per group), most of the statistical methods are highly sensitive about non-normality and there is chance to wrongly detect non-normality, despite having normal data. It is recommended that when sample size is sufficient, only on highly non-normal data, nonparametric method should be used otherwise corresponding parametric methods should be preferred.

Any question?

Any question?

THANKS YOU

Together We Will Do A Great Job !