



Data Analysis

Dr. Andrea Benedetti



○ Plan

- General thoughts on data analysis
- Data analysis
 - for RCTs
 - for Case Control studies
 - for Cohort studies
- Issues



Basic steps – Descriptive Stats

- Start with univariable descriptive statistics for each variable
 - Understand the distribution of your variables
 - Histograms
 - Missing values?
 - Data quality checks (e.g. reasonable values for height, weight, etc.)
 - Mins, Maxes
 - Logical consistency
 - What units?
- Bivariable descriptive statistics
 - cross tabulations
 - scatterplots
- Descriptives should tell most of the story most of the time



Basic steps -- Modelling

- Modelling (regression) should be a small part of total time spent
- Descriptives should tell you everything you need to know for models
- Minimize total number of models run
 - Decreases false positives
 - Increases reproducibility



RCTS



RCTs

- Randomization (usually) takes care of confounding
- Two types of analyses:
 - Pre-specified in the protocol
 - Findings form the basis for guidelines, etc.
 - Generally: intention to treat, unadjusted, in the whole population
 - Secondary analyses
 - may or may not be prespecified
 - more exploratory in nature

RCTs – Which participants should be analyzed?

- Intention to treat
 - ITT: Analyze everyone as they were randomized, even if...
 - they did not comply with treatment
 - they were ineligible
 - they were lost to follow up
 - Maintains the benefits of randomization
 - most valid
 - May underestimate the treatment effect
 - Most conservative



Per protocol analysis

- definitions vary– make sure you specify!
- subjects have now self selected into treatment groups
 - especially 'as treated'
 - must adjust for confounders!!

RCTs – Adjust?

- Confounding
 - Main analysis *usually* does not consider adjusting
 - Should consider adjusting
 - when there is imbalance on important confounders
 - as a sensitivity analysis
 - when using a per protocol analysis
 - if there is variable follow up time
 - When adjustment will be used, and for which variables should be pre-specified



RCTs -- Subgroups

- Subgroup analyses
 - Define which subgroups a priori
 - Use strict criterion
 - Interpret sceptically!! Especially for subgroup analyses not pre-specified



RCT...Primary analysis

What kind of outcome?

- Binary outcome
 - Chi square or Fisher exact test
- Continuous outcome
 - T-test
- Survival
 - Log rank test
- Counts
 - Test for counts

The Logic is Always the Same:

1. Assume nothing is going on (assume H_0)
2. Calculate a test statistic (Chi-square, t)
3. How often would you get a value this large for the test statistic when H_0 is true? (In other words, calculate p)
4. If $p < .05$, reject H_0 and conclude that something is going on (H_A)
5. If $p > .05$, do not conclude anything.



T-test for two means

- Is the difference in means that we observe between two groups more than we'd expect to see based on chance alone?
- $H_0: \mu_1 = \mu_2$
- $H_A: \mu_1 \neq \mu_2$

T-test, independent samples, pooled variance

If you assume that the standard deviation of the characteristic is the same in both groups, you can pool all the data to estimate a common standard deviation. This maximizes your degrees of freedom (and thus your power).

pooling variances:

$$s_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2}{n-1} \quad \text{and} \quad (n-1)s_x^2 = \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

$$s_y^2 = \frac{\sum_{i=1}^m (y_i - \bar{y}_m)^2}{m-1} \quad \text{and} \quad (m-1)s_y^2 = \sum_{i=1}^m (y_i - \bar{y}_m)^2$$

$$\therefore s_p^2 = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2 + \sum_{i=1}^m (y_i - \bar{y}_m)^2}{n+m-2}$$

$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

Degrees of Freedom!

T-test, pooled variances

$$T = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{m}}} \sim t_{n+m-2}$$

$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

χ^2 Test of Independence for Proportions

- Does a relationship exist between 2 categorical variables?
- Assumptions
 - Multinomial experiment
 - All expected counts ≥ 5
- Uses two-way contingency table

χ^2 Test of Independence

Hypotheses & Statistic

Hypotheses:

H₀: Variables Are Independent

H_A: Variables Are Related (Dependent)

Test Statistic:

$$\chi^2 = \sum_{\text{all cells}} \frac{[n_{ij} - \hat{E}(n_{ij})]^2}{\hat{E}(n_{ij})}$$

Observed count (points to n_{ij})

Expected count (points to $\hat{E}(n_{ij})$)

Degrees of freedom: $(r - 1)(c - 1)$



Expected Count Calculation

$$\text{Expected count} = \frac{(\text{Row total}) * (\text{Column total})}{\text{Sample size}}$$

χ^2 Test of Independence

Example on HIV

You randomly sample **286** sexually active individuals and collect information on their HIV status and History of STDs. At the **.05** level, is there evidence of a **relationship**?

STDs Hx	HIV		Total
	No	Yes	
No	84	32	116
Yes	48	122	170
Total	132	154	286

χ^2 Test of Independence Solution

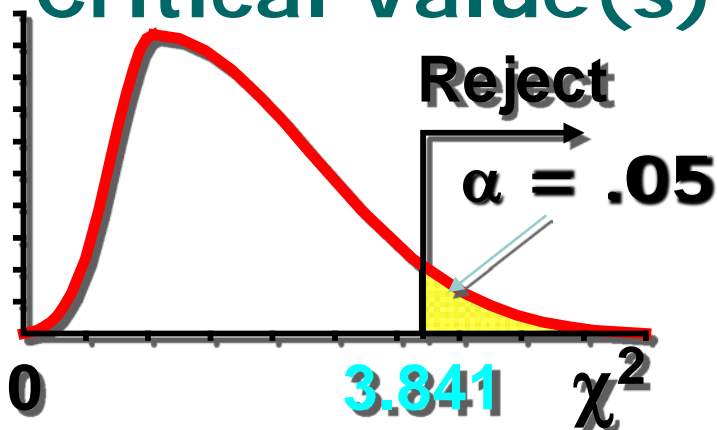
H_0 : No Relationship

H_A : Relationship

$\alpha = .05$

$df = (2 - 1)(2 - 1) = 1$

Critical Value(s):



Test Statistic:

$$\chi^2 = 54.29$$

Decision:

Reject at $\alpha = .05$

Conclusion:

There is evidence of a relationship



RCT... In secondary analysis

- consider adjusting for confounders, especially if 'Table 1' shows imbalance
 - use linear, logistic, Poisson or Cox regression depending on outcome type
 - we will see as we move along
- subgroups
- per protocol analyses

Example

Table 2. Death Rates and Hazard Ratios, Stratified According to CD4+ Cell Count.

CD4+ Count	Integrated Therapy				Sequential Therapy				Hazard Ratio (95% CI)*	P Value
	No. of Patients	No. of Person-Yr	No. of Deaths	Death Rate/100 Person-Yr (95% CI)	No. of Patients	No. of Person-Yr	No. of Deaths	Death Rate/100 Person-Yr (95% CI)		
All patients	429	467	25	5.4 (3.5–7.9)	213	223	27	12.1 (8.0–17.7)	0.44 (0.25–0.79)	0.003
≤200 cells/mm ³	273	281	23	8.2 (5.2–12.3)	138	137	21	15.3 (9.6–23.5)	0.54 (0.30–0.98)	0.04
>200 cells/mm ³	156	186	2	1.1 (0.1–3.9)	75	86	6	7.0 (2.6–15.3)	0.16 (0.03–0.79)	0.02

* Hazard ratios are for the integrated-therapy group, as compared with the sequential-therapy group.

- “After adjustment for baseline WHO status of HIV infection (stage 4 vs. stage 3), CD4+ cell count, age, sex, history of TB, extrapulmonary TB, and baseline HIV RNA level, the hazard ratio was 0.43 (95% CI, 0.25 to 0.77; P = 0.004).”
- “There was no interaction between the CD4+ count and the study groups (P = 0.57).”



OBSERVATIONAL STUDIES



Observational Studies

- Confounding!!
- Use a regression approach which allows
 - adjustment for confounders
 - investigation of effect modifiers

Linear regression: quick review

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

- attempt to fit a linear equation to observed data
 - One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable
 - Can include more than one variable
 - Estimated via maximum likelihood



Assumptions

- $Y|X \sim \text{Normal}$
- X-Y association is linear
- homoscedasticity
- independence

Interpreting parameters from simple linear regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

- Continuous X_1
 - β_1 is the expected change in Y for a 1 unit change in X_1
 - β_0 is the average Y when $X_1 = 0$
 - may or may not be logical!
- Binary X_1 (e.g. gender)
 - β_1 is the average difference in Y for subjects with $X_1 = 1$ vs. $X_1 = 0$
 - β_0 is the average Y when $X_1 = 0$

Interpreting parameters from multiple linear regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

- β_1 is the expected change in Y for a 1 unit change in X1 for subjects with the same value for X2
 - we might say β_1 is the expected change in Y adjusting for X2
- β_0 is the average Y when $X_1=0$ and $X_2=0$
 - may or may not be logical!

Adding an interaction term

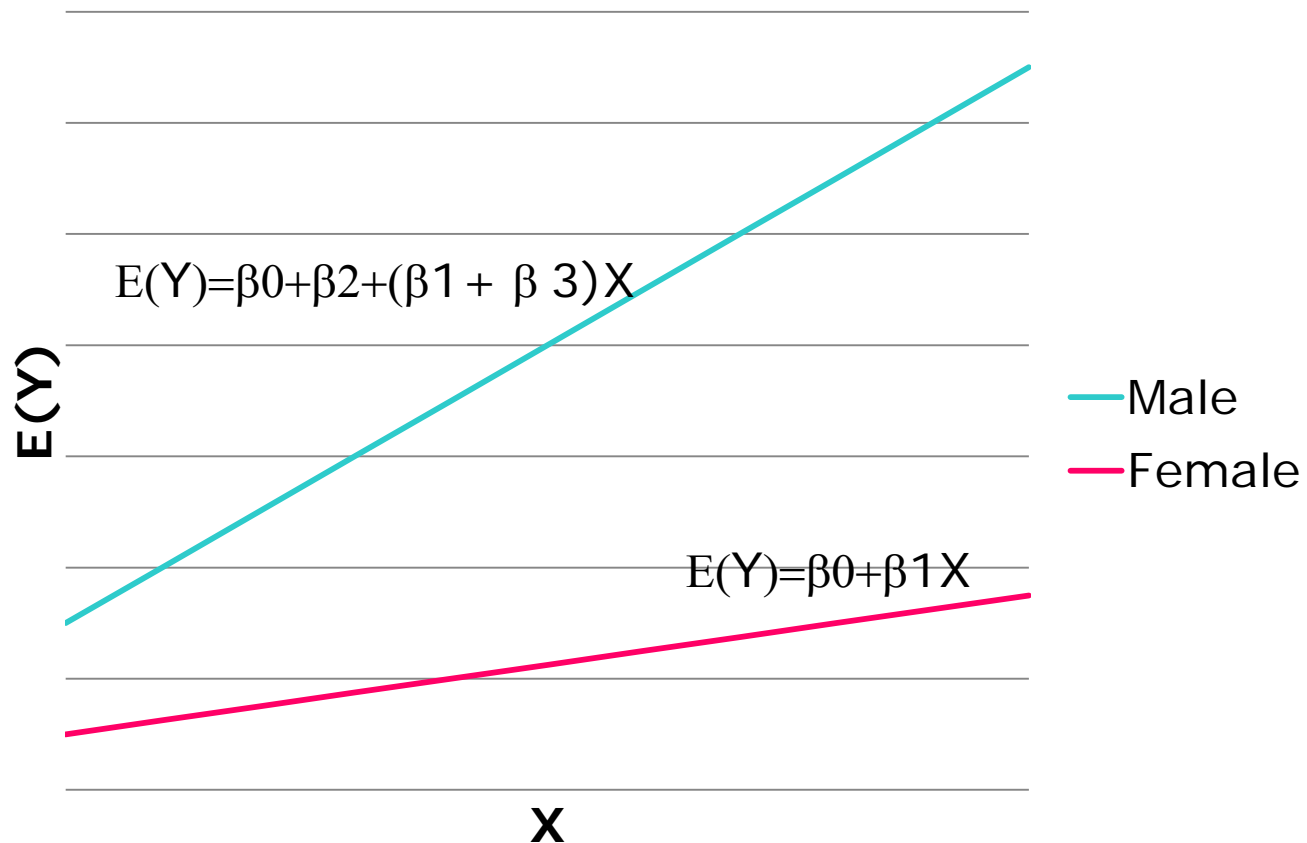
- Suppose we are interested in seeing whether there is effect modification by gender

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 \text{Male}_i + \beta_3 \text{Male}_i * X_1 + \varepsilon_i$$

- β_1 is the expected change in Y for a 1 unit change in X1 among Females
- β_2 is the average difference in Y between Males and Females when $X_1 = 0$
- $\beta_1 + \beta_3$ is the expected change in Y for a 1 unit change in X1 among Males

Adding an interaction term

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 \text{Male}_i + \beta_3 \text{Male}_i * X_1 + \varepsilon_i$$





ANALYZING DATA FROM CASE CONTROL STUDIES

Case Control Studies

- Sampling is based on disease status
 - Then exposure is ascertained
- The usual parameter of interest is the odds ratio

- $OR = (\text{odds of } E+ \text{ in the } D+) / (\text{odds of } E+ \text{ in the } D-)$
 $= (\text{odds of } D+ \text{ in the } E+) / (\text{odds of } D- \text{ in the } E-)$

$$= (A/B) / (C/D)$$

$$= AD/BC$$

		Disease	
		+	-
Exposure	+	A	B
	-	C	D

Case control studies

- Typically use logistic regression
 - Extends linear regression to deal with a binary outcome variable
 - Outcome variable is binary $Y \sim \text{Binomial}(p)$
 - Adjust for important confounders
 - Consider pre-specified interactions
- Logistic regression estimates the probability of an event occurring

$$p = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots}}$$

- $\text{logit}(p) = \log_e(p/(1-p))$
- $\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$

Interpreting parameters from a logistic regression

- Consider:
 - $\text{logit}(Y_{\text{exp}}) = \beta_0 + \beta_1 X_1$
- Among the exposed
 - $\text{logit}(Y_{\text{exp}}) = \beta_0 + \beta_1 \dots \dots \dots (1)$
- Among the unexposed
 - $\text{logit}(Y_{\text{unexp}}) = \beta_0 \dots \dots \dots (2)$
- Subtract (2) from (1):
 - $\text{logit}(Y_{\text{exp}}) - \text{logit}(Y_{\text{unexp}}) = \beta_1$

So...

$$\log\left(\frac{Y_{\text{exp}}}{1 - Y_{\text{exp}}}\right) - \log\left(\frac{Y_{\text{unexp}}}{1 - Y_{\text{unexp}}}\right) = \beta_1$$

$$\log\left(\frac{Y_{\text{exp}}}{1 - Y_{\text{exp}}} \bigg/ \frac{Y_{\text{unexp}}}{1 - Y_{\text{unexp}}}\right) = \beta_1$$

$$\text{OR} = \exp(\beta_1)$$

Example logistic regression

Table 2 OR of *Mycobacterium tuberculosis* infection (MTI) by BCG vaccination and other demographic characteristics, among 953 children and young adults in East Greenland

Demographic characteristics	Total N	MTI N (%)	OR* (95% CI)		p Value†
			Unadjusted estimate	Adjusted‡ estimate	
All	953	280 (29)			
BCG-vaccinated					
No	181	103 (57)	1	1	
Yes	772	177 (23)	0.23 (0.16 to 0.32)	0.52 (0.32 to 0.85)	0.01
Sex					
Female	509	132 (26)	1	1	
Male	444	148 (33)	1.43 (1.08 to 1.89)	1.70 (1.24 to 2.33)	0.001
Ethnicity					
Inuit	915	276 (30)	1	1	
Non-Inuit	38	4 (11)	0.27 (0.10 to 0.78)	0.44 (0.14 to 1.38)	0.16
Place of residence					
Town	688	203 (30)	1	1	
Settlement	265	77 (30)	0.98 (0.72 to 1.34)	1.31 (0.90 to 1.90)	0.20

*ORs relate to the odds of being *M. tuberculosis* infected defined by QFT positivity.

†Test for homogeneity based on the analysis with adjustment.

‡Adjusted for BCG, age, sex and ethnicity. QTF, QuantiFERON.



Analyzing matched case-control studies

- Advantage of matching: helps to control confounding at the design stage
- What variables to match on?
 - Strong/hard to measure confounders
 - Not too many variable



Conditional logistic regression

- Must account for matching in the analysis
 - Matching induces a bias in the effect estimates
 - Otherwise effect estimates will be biased towards null



Conditional logistic regression

- Used when we have cases matched to controls
- We are interested in the comparison WITHIN strata
- Cannot estimate the effect of the matching factor
 - but can estimate the interaction between that factor and another variable

Example

All cases with hepatotoxicity

Hepatotoxicity of Pyrazinamide Cohort and Case-Control Analyses

Kwok C. Chang¹, Chi C. Leung¹, Wing W. Yew², Tat Y. Lau¹, and Cheuk M. Tam¹

¹Tuberculosis and Chest Service,
Grantham Hospital, Hospital Autl

**TABLE 2. UNIVARI
STARTING TREATN**

Explanatory Variables
Sex (matched)
Female
Male
Age, yr (matched), me
Range*

**TABLE 4. MULTIVARIABLE CONDITIONAL LOGISTIC REGRESSION
ANALYSIS OF HEPATOTOXICITY FROM 12 OR MORE WEEKS AFTER
STARTING TREATMENT INVOLVING 33 CASES AND 96 MATCHED
CONTROL SUBJECTS WITH NO PREVIOUS HEPATOTOXICITY**

Explanatory Variables	OR (95% CI)	P
Predominant regimens received within 4 weeks preceding hepatotoxicity, with reference to regimens comprising H and R		0.02
Pyrazinamide-containing regimens: HRZ or HZ or RZ	2.5 (1.2–5.5)	0.02
Other regimens	4.2 (1.3–13.3)	0.01
Hepatitis B	2.7 (1.2–6.1)	0.02
Hepatitis C	4.6 (1.1–20.1)	0.04

For definition of abbreviations, see Table 3.

Covariates included in the analysis were the same as those shown in the legend
of Table 3.

I Chest Unit,

P

1.00

0.97

pp 1391–1396, 2008



General approach for case control studies

- First, compare cases and controls on important covariates
- Use logistic regression, to estimate the exposure effect adjusted for confounders
 - If matching was used use conditional logistic regression
- Investigate effect modification
- Check assumptions



ANALYZING DATA FROM COHORT STUDIES



Analyzing data from cohort studies

- Subjects are recruited and followed over time to see if the outcome develops
- Several different approaches are possible depending on the type of outcome

Frame work of Cohort studies

		Disease Status			
		Total	Yes	No	
Exposure Status	Yes	<div><div><div>$a+b$</div><div>a</div></div><div>b</div></div>	Study cohort		
	No	<div><div><div>$c+d$</div><div>c</div></div><div>d</div></div>	Comparison cohort		
	<div><div>N</div><div>$a+c$</div><div>$b+d$</div></div>				

Overview of statistical approaches for cohort studies

TABLE 1. Overview of analytical methods for cohort studies

Outcome	Summary measure	Comparison		Measure of association
		Exposed/unexposed (2-sample)	Multiple (regression)	
Events in person-years	Incidence rate	$(O-E)^2/\text{var}$	Poisson	Relative incidence
Time to event	Kaplan-Meier/maximum likelihood estimates	Logrank or Mantel-Haenszel/likelihood ratio test	Proportional hazards/parametric	Relative hazard/relative percentile or time
Time to event; exposures changing	Extended Kaplan-Meier	Extended logrank	Proportional hazards, staggered entries	Relative hazard
Case in nested case-control	Proportion exposed	Paired chi-square or McNemar	Conditional logistic	Odds ratio
Case in nested case-cohort	Proportion exposed	(Robust) logrank	Proportional hazards, staggered entries	Relative hazard
Intermediate outcome repeatedly measured	Change		Regression for correlated data; marginal, conditional, random effects	Differences in change over time



Cox Proportional Hazards

- The outcome of interest is time to event



Characteristics of Cox Regression

- Does not require that you choose some particular probability model to represent survival times.... robust
- *Semi*-parametric
(Kaplan-Meier is *non-parametric*;
exponential and Weibull are *parametric*)
- Easy to incorporate time-dependent covariates—covariates that may change in value over the course of the observation period



Main Assumptions of Cox Regression

- Proportional hazards assumption: the hazard for any individual is a fixed proportion of the hazard for any other individual
- Multiplicative risk
- Independent events
- Uninformative censoring

Recall: The Hazard function

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t / T \geq t)}{\Delta t}$$

In words: the probability that ***if you survive to t***, you will succumb to the event in the next instant.

The model

Components:

- A baseline hazard function that is left unspecified
- A linear function of covariates that is exponentiated. (=the hazard ratio)

$$h_i(t) = \boxed{\lambda_0(t)} e^{\beta_1 x_{i1} + \dots + \beta_k x_{ik}}$$

Can take on any form!

$$\log h_i(t) = \boxed{\log \lambda_0(t) + \beta_1 x_{i1} + \dots + \beta_k x_{ik}}$$

The model: binary predictor

$$HR_{lung\ cancer / smoking} = \frac{h_i(t)}{h_j(t)} = \frac{\cancel{\lambda_0(t)} e^{\beta_{smoking}(1) + \cancel{\beta_{age}(60)}}}{\cancel{\lambda_0(t)} e^{\beta_{smoking}(0) + \cancel{\beta_{age}(60)}}} = e^{\beta_{smoking}(1-0)}$$

$$HR_{lung\ cancer / smoking} = e^{\beta_{smoking}}$$

This is the hazard ratio for smoking adjusted for age.

Example: Cox

Table 2. Death Rates and Hazard Ratios, Stratified According to CD4+ Cell Count.

CD4+ Count	Integrated Therapy				Sequential Therapy				Hazard Ratio (95% CI)*	P Value
	No. of Patients	No. of Person- Yr	No. of Deaths	Death Rate/ 100 Person-Yr (95% CI)	No. of Patients	No. of Person- Yr	No. of Deaths	Death Rate/ 100 Person-Yr (95% CI)		
All patients	429	467	25	5.4 (3.5–7.9)	213	223	27	12.1 (8.0–17.7)	0.44 (0.25–0.79)	0.003
≤200 cells/mm ³	273	281	23	8.2 (5.2–12.3)	138	137	21	15.3 (9.6–23.5)	0.54 (0.30–0.98)	0.04
>200 cells/mm ³	156	186	2	1.1 (0.1–3.9)	75	86	6	7.0 (2.6–15.3)	0.16 (0.03–0.79)	0.02

* Hazard ratios are for the integrated-therapy group, as compared with the sequential-therapy group.

- “After adjustment for baseline WHO status of HIV infection (stage 4 vs. stage 3), CD4+ cell count, age, sex, history of TB, extrapulmonary TB, and baseline HIV RNA level, the hazard ratio was 0.43 (95% CI, 0.25 to 0.77; P = 0.004).”
- “There was no interaction between the CD4+ count and the study groups (P = 0.57).”



Time-dependent covariates

- Covariate values for an individual may change over time
 - E.g. If you are evaluating the effect of weight on diabetes risk over a long study period, subjects may gain and lose large amounts of weight, making their baseline weight a less than ideal predictor.
- Cox regression can handle these time-dependent covariates!



Time-dependent covariates

- Ways to look at drug use:
- Not time-dependent
 - Ever/never during the study
 - Yes/no use at baseline
 - Total months use during the study
- Time-dependent
 - Using drug use at event time t (yes/no)
 - Months of drug use up to time t



Overview of cohort study data analysis

- First compare exposed and not exposed subjects on important covariates
- Use a regression model to adjust the exposure effect for important covariates
 - Logistic, poisson, cox depending on interest/how study was conducted
- Consider effect modification
- Consider time dependent covariates
- Check assumptions



**NO MATTER WHAT METHOD
YOU USE...**



Your statistical methods....

- Should match your objectives
- Should be appropriate for the type of outcome you have
- Should be appropriate for the study design you have chosen

Functional form

Think about what form the association

Downloaded from thorax.bmj.com on July 10, 2014 - Published by group.bmj.com

Thorax Online First, published on June 26, 2014 as 10.1136/thoraxjnl-2014-205688

Tuberculosis

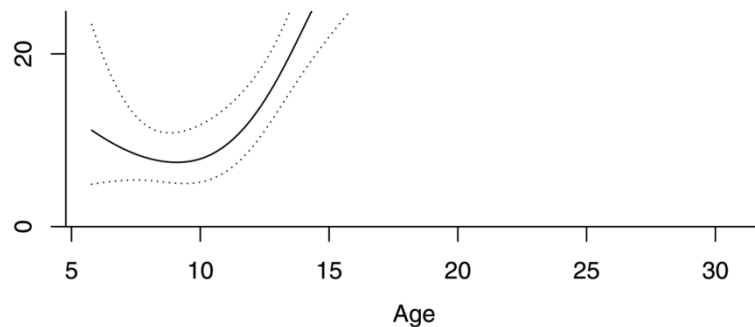
ORIGINAL ARTICLE

The effectiveness of BCG vaccination in preventing *Mycobacterium tuberculosis* infection and disease prevalence of MTI by age. Age was modelled using a

S W Michelsen,¹ B Soborg,¹ A Koch,¹ L Carstensen,¹ S Møller,² E M Jorgensen,² T Lillebaek,³ H C F Sorensen,⁴ J Wohlfahrt,¹ M Melbye¹

spline.

Predicted MTI prevalence (%)



ng approach

|?

Which variables are important?

- Subject matter drives this to a large extent
 - Known confounders
 - Eliminate intermediates
 - a variable in the causal pathway between the exposure and the outcome
- Directed acyclic graphs (DAGs)?
- Statistical significance?
- Data driven selection
 - e.g. stepwise selection



Missing data

- Think carefully about how to address it
 - Drop individuals?
 - Drop variables?
 - Multiple imputation?
 - Usually the best option.



Outliers

- Outliers can have a lot of influence on your statistical tests
- Outliers may be mistakes, or could be true data
 - Should be examined to ensure they are true data
 - If erroneous, could be removed or corrected

Example: interaction

Detecting Tuberculosis Infection in HIV-infected Children: A Study of Diagnostic Accuracy, Confounding and Interaction

Anna M. Mandalakas, MD,†‡ Susan van Wyk, MD,‡ H. Lester Kirchner, PhD,§ Gerhard Walzl, MD, PhD,¶
Mark Cotton, MD, PhD,|| Helena Rabie, MD,|| Belinda Kriel, NHD Med Tech,¶ Robert P. Gie, FCP,‡
H. Simon Schaaf, MD, PhD,‡ and Anneke C. Hesseling, MD, PhD‡*

Background: Accurate identification of *Mycobacterium tuberculosis* infection in young and HIV-infected children could guide delivery of preventive therapy, improve resource utilization and help prevent tuberculosis.

Key Words: tuberculosis, HIV, latent tuberculosis infection, pediatrics, interferon- γ release assays, tuberculin skin test

(*Pediatr Infect Dis J* 2013;32: e111–e118)

- Cohort study
- Logistic regression assessed the association among test positivity, age, nutritional and HIV status, while controlling for *M. tuberculosis* exposure, bacille Calmette–Guérin vaccination and prior tuberculosis treatment.

Logistic regression example

- Logistic regression adjusted for age, prior BCG vaccination, prior TB treatment, chronic malnutrition status and HIV status, and included interaction for HIV status and age.
- Significant interaction between age and HIV status ($P = 0.0052$, $P = 0.0404$, respectively).

Covariates	TST* OR (95% CI)	
	Unadjusted	Adjusted [†]
	(n = 247)	(n = 247)
TB contact score	1.18 (1.04, 1.35)	1.14 (0.99, 1.30)
BCG vaccination		0.72 (0.21, 2.46)
Prior TB treatment		0.69 (0.32, 1.50)
Age (yr) effect [‡]		
HIV infected		1.04 (0.93, 1.16)
HIV uninfected		1.23 (1.08, 1.40)
HAZ score effect [§]		1.08 (0.94, 1.25)
HIV infected		
HIV uninfected		
HIV effect [‡]		
25%tile (1.6 yr)		0.50 (0.27, 1.17)
Median (3.3 yr)		0.75 (0.41, 1.39)
75%tile (6.9 yr)		1.41 (0.67, 2.94)



Other complications

- Clustered data
 - longitudinal data
 - families, households, etc
- Observations are not independent
- Must account for the correlation between observations on the same person/in the same family/etc.
- Mixed models, or marginal models estimated via GEE



Checking assumptions

- Important, but rarely presented
- Must check that the assumptions of the model are met!
 - Model fit (predicted vs. observed)
 - Outliers
 - Influential observations
 - Key – is any one observation “driving” the results?



Wrapping up – Data analysis

- Start with descriptives
 - should tell most of the story most of the time
 - should inform the next step (modelling)
- For RCTs
 - Primary analysis is usually ITT, unadjusted, simple test of the outcome of interest
 - What test depends on the type of outcome
 - Secondary analyses may include adjusting for confounders, per protocol, subgroups
 - use regression appropriate for the type of outcome



Wrapping up – Data analysis 2

- For observational studies, the primary analysis will usually need to adjust for confounders
 - Use regression methods appropriate for the outcome and the study design
 - logistic regression for case control studies
 - conditional logistic regression for matched case control studies
 - Cox, Poisson, or logistic for cohort studies



Wrapping up – Data analysis 3

- In all cases, consider:
 - Functional form of exposure and covariates in regression
 - How to choose confounders to include in the regression
- Missing data
- Checking assumptions



Software

- R is available on the web
 - <http://cran.r-project.org/>
 - Free
 - Flexible
 - Lots of online training resources
 - User friendly?
- Stata, SAS
- SPSS, others

Binary or categorical outcomes (proportions)

Outcome Variable	Are the observations correlated?		Alternative to the chi-square test if sparse cells:
	independent	correlated	
Binary or categorical (e.g. fracture, yes/no)	<p>Chi-square test: compares proportions between two or more groups</p> <p>Relative risks: odds ratios or risk ratios</p> <p>Logistic regression: multivariable technique used when outcome is binary; gives multivariable-adjusted odds ratios</p>	<p>McNemar's chi-square test: compares binary outcome between correlated groups (e.g., before and after)</p> <p>Conditional logistic regression: multivariable regression technique for a binary outcome when groups are correlated (e.g., matched data)</p> <p>Mixed models/GEE modeling: multivariate regression technique for a binary outcome when groups are correlated (e.g., repeated measures)</p>	<p>Fisher's exact test: compares proportions between independent groups when there are sparse data (some cells <5).</p> <p>McNemar's exact test: compares proportions between correlated groups when there are sparse data (some cells <5).</p>

Continuous outcome (means)

Outcome Variable	Are the observations independent or correlated?		Alternatives if the normality assumption is violated (and small sample size):
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	<p>Ttest: compares means between two independent groups</p> <p>ANOVA: compares means between more than two independent groups</p> <p>Pearson's correlation coefficient (linear correlation): shows linear correlation between two continuous variables</p> <p>Linear regression: multivariable regression technique used when the outcome is continuous; gives slopes</p>	<p>Paired ttest: compares means between two related groups (e.g., the same subjects before and after)</p> <p>Repeated-measures ANOVA: compares changes over time in the means of two or more groups (repeated measurements)</p> <p>Mixed models/GEE modeling: multivariate regression techniques to compare changes over time between two or more groups; gives rate of change over time</p>	<p><u>Non-parametric statistics</u></p> <p>Wilcoxon sign-rank test: non-parametric alternative to the paired ttest</p> <p>Wilcoxon sum-rank test (=Mann-Whitney U test): non-parametric alternative to the ttest</p> <p>Kruskal-Wallis test: non-parametric alternative to ANOVA</p> <p>Spearman rank correlation coefficient: non-parametric alternative to Pearson's correlation coefficient</p>



EXTRAS



Nonparametric vs. Parametric

- Parametric tests –
 - make assumptions about the distributions of our variables that may or may not be true
- Nonparametric tests avoid those assumptions
 - are usually based on ranking
 - are usually less powerful

Poisson Regression

- Appropriate when subjects are followed for varying lengths of time
- Outcome is a count
- Similar to logistic regression, however now $Y \sim \text{Poisson}$
- The link function is log
$$\log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$$
$$\exp(\beta_1) = \text{RR}$$