THE GEORGE INSTITUTE
for International Health

APBG Meeting
Sydney, Australia – 22/08/2012

Relative risk estimation in prospective studies: alternatives to logistic regression

Laurent Billot, MSc DEA AStat
Director, Statistics and Data Management
The George Institute for Global Health

Why this talk?

- Odds ratios ≠ risk ratios (common event)
- Misinterpretations
- No need to use ORs for cross-sectional or prospective studies
- Alternatives to logistic regression exist

Outline

- Definitions and interpretations
- Simple examples
- When ORs are a problem
- Logistic regression
- Alternative methods
- Example: the ADVANCE study
- Concluding remarks and recommendations

DEFINITIONS AND EXAMPLES

Definition: risks and risk ratios

Exposed to risk factor?	Have disease?		
	Yes	No	Total
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	n

- risk of disease in exposed subjects = a / (a+b)
- risk in non-exposed subjects = c / (c+d)
- Risk ratio (exposed / non-exposed) = {a / (a+b)} / {c / (c+d) }

Definition: odds and odds ratios

Exposed to risk factor?	Have disease?		
	Yes	No	Total
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	n

- odds in exposed subjects = a / b
- odds in non-exposed subjects = c / d
- odds ratio = { a / b } / { c / d } = { a × d } / { b × c }

Risk and risk ratio: interpretation

- Risk in (non) exposed subjects:
  - Proportion of subjects who developed disease among all (non) exposed subjects
    Or
  - Probability of developing disease in (non) exposed subjects

- Example (fictitious):
  - Risk of stroke in males = 20%
    ➔ males have a 20% chance of having a stroke
  - Risk of stroke females = 10%
    ➔ females have a 10% chance of having a stroke
  - Relative risk = 2
    ➔ Males are twice as likely to have a stroke than females

Odds and odds ratio: interpretation

- Odds in (non) exposed subjects:
  - Relative probability of developing disease compared with not developing disease among (non) exposed subjects
    Or
  - Ratio of patients developing disease to patients not developing disease in (non) exposed subjects

- Example:
  - Odds in males = 1 to 4 (0.25)
    ➔ for every male experiencing a stroke, 4 will not have one
  - Odds in females = 1 to 9 (0.1111)
  - ➔ for every female experiencing a stroke, 9 will not have one
  - Odds-ratio = 2.25
    ➔ The odds of stroke is 2.25 times greater in males subjects

Odds ratio vs. risk ratio: real example

Smoker at entry?	Cardiovascular death during follow-up?		
	Yes	No	Total
Yes	31	1386	1417
No	15	1883	1898
Total	46	3269	3315

Source: EGAT study

- risk of death in smokers = 31/1417 = 0.0219 (=2.2%)
- risk in non-smokers = 15/1898 = 0.0079 (=0.8%)
- Relative risk (smoker / non-smoker) = 0.0219 / 0.0079 = 2.77

- odds in smokers = 31/1386 = 0.0224 (= 1/44.7)
- odds in non-smokers = 15/1883 = 0.0080 (=1/125.5)
- odds ratio = 0.0224 / 0.0080 = 2.81

Odds ratio vs. relative risk: modified example

Smoker at entry?	Cardiovascular death during follow-up?		
	Yes	No	Total
Yes	620	797	1417
No	300	1598	1898
Total	920	2395	3315

modified EGAT data

- risk of death in smokers = 620/1417 = 0.4375
- risk in non-smokers = 300/1898 = 0.1581
- Relative risk (smoker / non-smoker) = 0.4375 / 0.1581 = 2.77

- odds in smokers = 620/797 = 0.7779
- odds in non-smokers = 300/1598 = 0.1877
- odds ratio = 0.7779 / 0.1877 = 4.14

Emma's example

Diagnosis?	BP treatment prescribed?		
	Yes	No	Total
MI	357	47	404
Stroke	206	148	354
Total	563	195	758

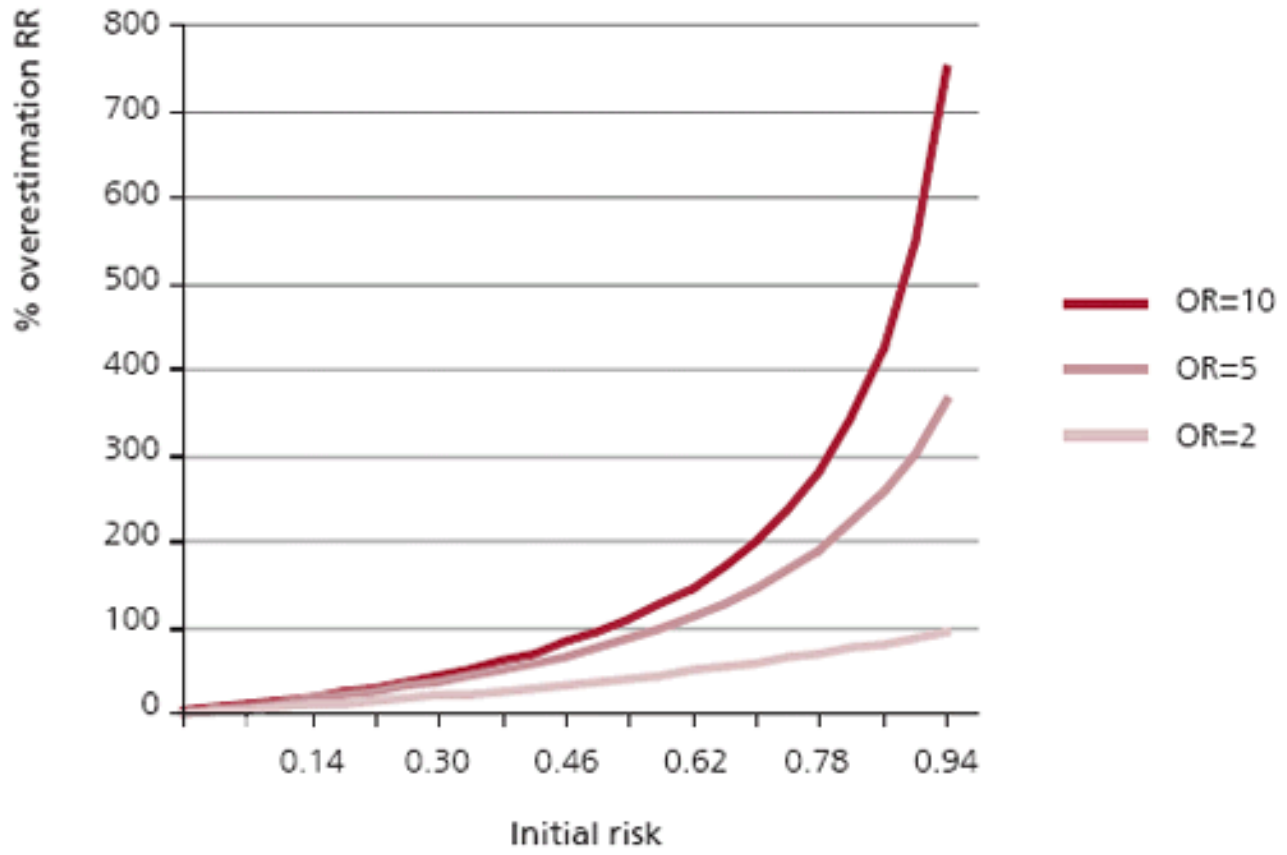
- Outcome: getting blood pressure treatment after cardiovascular event
- Overall incidence = 74%
- Relative risk (MI / Stroke) = 1.52
- Odds ratio = 5.46

Why use or not use odds ratio?

- Pros:
  - Good approximation of relative risk when event is rare
  - Only choice for case control studies
  - Good mathematical properties
  - Logistic regression
- Cons:
  - Difficult to understand odds and even more difficult to understand odds ratios
  - Often interpreted as relative risk (e.g. many examples of incorrect media reports)
  - Can be very different from relative risk if event is common (>10-15%)
  - Not needed in prospective or cross-sectional studies

OR vs RR as a function of prevalence

Figure. Relationship between the odds ratio (OR) and the % overestimation of the relative risk (RR) depending on initial risk



MODELS FOR BINARY DATA

Notations

- Outcome Y is a binary response (value 0 or 1)
- P is the probability that Y equals 1, or the proportion of 'positive' responses
- Odds = P / (1 – P)
- X_1, X_2, …, X_k are k covariates (e.g. gender, age, smoking status, etc.)

Logistic regression

- Most common model for binary data
- Models the log of the odds (logit link) as function of a linear combination of covariates (+ error ε)
- Assumes errors ε follow a binomial distribution

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = b_0 + b_1 x_1 + \cdots + b_k x_k + \varepsilon$$

Logistic regression (2)

- Widely used and available in every software
- Logit link has good properties (canonical link)
- Easy to implement in standard software
- Numerically stable

  but

- Provides odds ratios, not relative risks
- Discrepancies when outcome is common

Alternatives methods providing risk ratios

1. Simple transformations
2. Stratified Mantel-Haenszel test
3. **Log-binomial regression**
4. **Robust Poisson regression**
5. Robust Cox regression
6. Other regression-based techniques

Simple transformation

- When analysis is not adjusted, can derive risk ratio from odds ratio using following formula:

  $RR = OR / \{(1-P^0)+(OR*P^0)\}$

  where P^0 is the probability of event in the reference group

- Biased estimates when more than one predictor
- Difficult to get confidence intervals

McNutt et al 2003

Stratified Mantel-Haenszel procedure

- MH analysis calculates risk ratio within each level of a confounder (stratum)
- Stratum-specific risk ratios pooled to compute one adjusted risk ratio
- Easy to implement and unbiased

  but

- Does not handle multiple confounders unless nested
- Does not allow adjustment for continuous covariates

McNutt et al 2003

Log-binomial regression

- Uses log link instead of logit link used in logistic regression

$$\log(p) = b_0 + b_1 x_1 + \cdots + b_k x_k + \varepsilon$$

- Models probability of event rather than odds
- Directly provides risk ratios in prospective studies and ratios of prevalence in cross-sectional data
- Easy to implement in standard software
- Convergence issues

Wacholder 1986

Blizzard and Hosmer 2006

Poisson regression with robust variance

- Also uses log link but assumes Poisson distribution instead of binomial
- Standard errors too large with standard Poisson regression
- Need to correct variance using robust sandwich method
- Easy to implement in standard software
- Estimates not as efficient as those from log-binomial model
- Estimated probabilities can exceed 1

Zou 2004

Cox regression with robust variance

- Cox used to analyse events occurring at different times
- Calculates hazard ratios
- To make the hazard ratio equal to the relative risk:
  - set same event time for all observations
  - use Breslow method for handling ties
- Requires robust variance to obtain correct variance estimation
- Equivalent to the robust Poisson

Other regression-based techniques

- Can use other binary regression models
- For example complementary log-log regression
- Need to estimate standard errors using the delta method
- No direct software procedure

Penman and Johnson 2009

Cummings 2009

APPLICATION TO THE ADVANCE STUDY

ADVANCE study

- RCT
- 11,140 patients with Type 2 diabetes
- Picked 3 outcomes:
  1. Primary combined macrovascular and microvascular endpoint (common 19%)
  2. Dementia (rare event 1%)
  3. Hospitalisation (very common event 44%)
- Main independent variable: Glucose treatment (Intensive vs Standard glucose therapy)
- 6 other covariates: Age, Sex, Blood pressure treatment, history of macrovascular disease, history of microvascular disease, region

2x2 table (SAS proc freq)

Primary endpoint

Frequency Row Pct	Event	No event	Total
Intensive	1009 18.11	4562 81.89	5571
Standard	1116 20.04	4453 79.96	5569
Total	2125	9015	11140

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	0.8825	0.8028	0.9701
Cohort (Col1 Risk)	0.9038	0.8371	0.9758

SAS code

1. Logistic regression:

```
proc genmod data = ADVANCE ;
    class     TRTGLUC;
    model     PRIMARY_JOINT = TRTGLUC / dist = binomial link =
    logit;
    estimate 'Beta' TRTGLUC 1 -1/ exp;
run;
```

2. Log-binomial regression:

```
proc genmod data = ADVANCE ;
    class     TRTGLUC;
    model     PRIMARY_JOINT = TRTGLUC / dist = binomial link = log;
    estimate 'Beta' TRTGLUC 1 -1/ exp;
run;
```

3. Poisson regression with robust variance:

```
proc genmod data = ADVANCE ;
    class     TRTGLUC PATIENT_ID;
    model     PRIMARY_JOINT = TRTGLUC / dist = poisson link = log;
    repeated subject = PATIENT_ID / type = unstr;
    estimate 'Beta' TRTGLUC 1 -1/ exp;
run;
```

Stata code

1. Logistic regression:

```
glm PRIMARY_JOINT TRTGLUC, fam(bin) link(logit) eform
```

2. Log-binomial regression:

```
glm PRIMARY_JOINT TRTGLUC, fam(bin) link(log) eform
```

3. Poisson regression with robust variance:

```
glm PRIMARY_JOINT TRTGLUC, fam(poisson) link(log)
   robust eform
```

Results: primary cardiovascular outcome

	Crude			Adjusted		
Method	**OR/RR**	**95% CI**		**OR/RR**	**95% CI**	
2x2 table RR	0.9038	0.8371	0.9758	N/A	N/A	N/A
2x2 table OR	0.8825	0.8028	0.9701	N/A	N/A	N/A
Logistic	0.8825	0.8028	0.9701	0.8850	0.8034	0.9749
Log-binomial	0.9038	0.8371	0.9758	0.9113	0.8465	0.9811
Robust poisson	0.9038	0.8371	0.9758	0.9101	0.8444	0.9810

Results: dementia outcome

	Crude			Adjusted		
Method	**OR/RR**	**95% CI**		**OR/RR**	**95% CI**	
2x2 table RR	1.2704	0.8720	1.8509	N/A	N/A	N/A
2x2 table OR	1.2734	0.8708	1.8620	N/A	N/A	N/A
Logistic	1.2734	0.8708	1.8620	1.3244	0.9032	1.9419
Log-binomial	1.2704	0.8720	1.8509	N/A *	N/A *	N/A *
Robust poisson	1.2704	0.8720	1.8509	1.3176	0.9044	1.9196

* ERROR: The mean parameter is either invalid or at a limit of its range for some observations.

Results: hospitalisations

	Crude			Adjusted		
Method	**OR/RR**	**95% CI**		**OR/RR**	**95% CI**	
2x2 table RR	1.0500	1.0068	1.0951	N/A	N/A	N/A
2x2 table OR	1.0908	1.0121	1.1756	N/A	N/A	N/A
Logistic	1.0908	1.0121	1.1756	1.1038	1.0218	1.1924
Log-binomial	1.0500	1.0068	1.0951	1.0431*	1.0065*	1.0809*
Robust poisson	1.0500	1.0068	1.0951	1.0552	1.0130	1.0992

* WARNING: The relative Hessian convergence criterion of 0.0089116856 is greater than the limit of 0.0001. The convergence is questionable.

Convergence issues in log-binomial model

- Unlike the logistic model, the log-binomial model places restrictions on the parameter space, and the maximum likelihood estimate (MLE) might occur on the boundary of the parameter space, in which case the algorithm will not converge
- More likely to occur with many covariates, especially continuous ones. Increases chance of a risk close to boundaries 0 or 1 in one “category”.

Possible solutions

- Change starting values for parameters: e.g. solutions from robust Poisson regression and/or a negative intercept (e.g. -10)
- Use “COPY” method (Deddens et al., 2003) : combine c-1 copies of the original data set with one copy where all values of the outcome variable are interchanged -> use weights instead
  - Same idea as adding a small constant to cell counts to remove the problem of zero counts when analysing categorical data
  - Need to multiple estimated standard error by square root of c to calculate 95% Cis
- Truncation methods (Wacholder 1986)

Example using COPY method

TABLE 1. Results for the Greenland data (4)

Method	Receptor	Stage2	Stage3
MLE* (original data)			
Estimate	1.5583	2.5382	5.8680
95% CI*	1.0487, 2.3155	1.1734, 5.4903	2.7458, 12.5406
Poisson approximation			
Estimate	1.6308	2.5207	5.9134
95% CI	1.0745, 2.4751	1.1663, 5.4479	2.7777, 17.5890
MLE (modified data), $c = 1{,}000$			
Estimate	1.5567	2.5235	5.8237
95% CI	1.0479, 2.3126	1.1699, 5.4432	2.7326, 12.4117
MLE (modified data), $c = 10{,}000$			
Estimate	1.5582	2.5367	5.8636
95% CI	1.0487, 2.3152	1.1730, 5.4855	2.7445, 12.5276

* MLE, maximum likelihood estimator; CI, confidence interval.

Petersen and Deddens 2006

SAS code: COPY method

```
data  COPY;
 set  ADVANCE (in=ORIGIN)
      ADVANCE (in=CHANGED);
      if     ORIGIN then WEIGHT=9999;
      if     CHANGED
      then  do;
            WEIGHT    = 1;
            DEMENTIA = 1 - DEMENTIA;
            HOSP      = 1 - HOSP;
      end;
run;

proc genmod data = COPY ;
   class    TRTGLUC SEX TRTBP HISTORY_MICRO
            HISTORY_MACRO REGION_NAME;
   model    DEMENTIA =  TRTGLUC AGE SEX TRTBP HISTORY_MICRO
                        HISTORY_MACRO REGION_NAME
                        / dist = binomial link = log;
   estimate 'Beta' TRTGLUC 1 -1/ exp;
   freq     WEIGHT;
run;
```

SAS code: Intercept method

```
proc genmod data = ADVANCE;
   class     TRTGLUC SEX TRTBP HISTORY_MICRO
             HISTORY_MACRO REGION_NAME;
   model     DEMENTIA =  TRTGLUC AGE SEX TRTBP HISTORY_MICRO
                         HISTORY_MACRO REGION_NAME
                         / dist = binomial link = log intercept=-10;
   estimate 'Beta' TRTGLUC 1 -1/ exp;
run;
```

Results: dementia and hospitalisations

	Dementia			Hospitalisations		
Method	**OR/RR**	**95% CI**		**OR/RR**	**95% CI**	
Log-binomial	N/A	N/A	N/A	1.0431[1]	1.0065	1.0809
Log-binomial (COPY)	1.2221[2]	0.8852	1.6873	1.0468[3]	1.0065	1.0886
Log-binomial (Int=-10)	1.3189	0.9070	1.9180	1.0689[1]	1.0101	1.1312
Log-binomial (COPY+Int)	1.3148[3]	0.9060	1.9081	1.0468[3]	1.0065	1.0886
Robust poisson	1.3176	0.9044	1.9196	1.0552	1.0130	1.0992

1. Hessian convergence criterion not met
2. Limited to c=300
3. c=10,000

Concluding remarks

- OR valid measure but misleading if interpreted as risk ratio, especially when outcome not rare (>10-15%)
- Available alternatives when relative risk (and not odds ratio) is the parameter of interest
- Consider log-binomial rather than logistic in RCT
- Can have convergence issues if too many counfounders (cross-sectional study)
- Use COPY or small intercept method
- Check results with robust Poisson regression.
- Use blind review to finalise choice of model

References

- General
  - **Lumley, T., R. Kronmal, and S. Ma. 2006. Relative risk regression in medical research: Models, contrasts, estimators, and algorithms. Working Paper 293, UW Biostatistics Working Paper Series. http://www.bepress.com/uwbiostat/paper293**
  - Petersen MR, Deddens JA. Re: ‘‘Easy SAS calculations for risk or prevalence ratios and differences.’’ (Letter). Am J Epidemiol 2006;163:1158–9.
  - Zou G. A Modified Poisson Regression Approach to Prospective Studies with Binary Data. Am J Epidemiol 2004; 159(7):702-6.
  - Blizzard, L. and Hosmer D. W. (2006). Parameter estimation and goodness-of-fit in log binomial regression. Biometrical Journal 48, 5–22.
  - Barros AJ, Hirakata VN. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. BMC Med Res Methodol. 2003 Oct 20;3(1):21.
  - McNutt L-A, Wu C, Xue X, Hafner JP (2003) Estimating relative risk in cohort studies and clinical trials of common events. American Journal of Epidemiology. 157: 940-943.
  - Penman, A.D., Johnson, W.D. Complementary log-log regression for the estimation of covariate-adjusted prevalence ratios in the analysis of data from cross-sectional studies (2009) Biometrical Journal, 51 (3), pp. 433-442.

References

- SAS
  - Spiegelman D, Hertzmark E. Easy SAS calculations for risk or prevalence ratios and differences. Am J Epidemiol 2005;162:199–200
  - Deddens & Petersen. (2003) Estimation of prevalence ratios when PROC GENMOD does not converge. Proceedings of the 28th Annual SAS Users Group International Conference, paper 270—28. Cary NC, SAS Institute Inc
  - http://www.ats.ucla.edu/stat/sas/faq/relative_risk.htm

- Stata
  - **Cummings, P. Methods for estimating adjusted risk ratios (2009) Stata Journal, 9 (2), pp. 175-196.**
  - http://www.ats.ucla.edu/stat/stata/faq/relative_risk.htm

Thank you