Estimating risk ratio from any standard design by doubling the cases

Yilin Ning, PhD

Centre for Quantitative Medicine, Duke-NUS Medical School

6 Jul 2022

Background

Risk ratio (RR) has a more intuitive interpretation than odds ratio (OR).

OR is more commonly reported in epidemiological and clinical studies.

- OR is estimated using the familiar logistic regression model.
- RR is estimated using the less familiar log-binomial/Poisson regression model.

Unlike OR, RR is not available from traditional analysis of case-control data.

- Does this restrict us to reporting OR for binary outcomes?

Background

- The doubling-of-cases approach[1] enables valid estimation of RR from cohort data, by applying the logistic regression to *a modified data set*.

- Application of this approach has been limited:
  - robust standard error (SE) required to account for the data modification,
  - not available in any statistical software package.

- A recent work[2] extended the application to case-control studies and implement the approach (for any design) as an R package[3].

1. Schouten et al. Risk Ratio and Rate Ratio Estimation in Case-Cohort Designs: Hypertension and Cardiovascular Mortality. *Stat. Med.* 1993;12(18):1733–1745.
2. Ning et al. Estimating risk ratio from any standard epidemiological design by doubling the cases. *BMC Medical Research Methodology.* 2022;22:157.
3. https://github.com/nyilin/DoublingOfCases

Objective

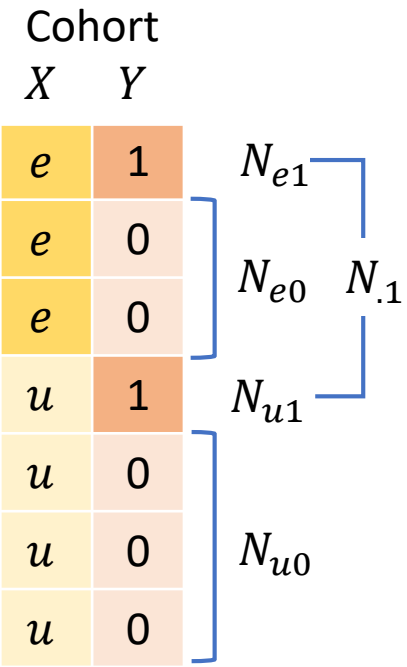
This session aims to provide:

- An intuitive introduction to the doubling of cases approach.
- A method to estimate RR from cohort and case-control data.
- The flexibility of reporting OR or RR (or both) regardless of study design.

Doubling of cases in cohort studies

Doubling of cases in cohort studies

Consider a cohort of N subjects with a binary outcome ($Y = 0, 1$) and binary exposure ($X = e, u$).



Calculation of the crude RR:

Original	$Y = 1$	$Y = 0$	Total	Prevalence	Crude RR
$X = e$	N_{e1}	N_{e0}	$N_{e.} = N_{e1} + N_{e0}$	$p_e = N_{e1}/N_{e.}$	$RR = p_e/p_u$
$X = u$	N_{u1}	N_{u0}	$N_{u.} = N_{u1} + N_{u0}$	$p_u = N_{u1}/N_{u.}$	

Doubling of cases in cohort studies (ctd.)

Now modify data by creating an additional record for each case,

but with the outcome changed to 0.

The *"expanded"* cohort, with outcome Y^* has $N + N_{.1}$ records,

$N_{.1}$ records as before with $Y^* = 1$ and N with outcome $Y^* = 0$.

Cohort

X	Y
e	1
e	0
e	0
u	1
u	0
u	0
u	0

Expanded

Cohort

X	Y^*
e	1
e	0
e	0
e	0
u	1
u	0
u	0
u	0
u	0

Original	$Y = 1$	$Y = 0$	**Total**	**Prevalence**	**Crude RR**
$X = e$	N_{e1}	N_{e0}	$N_{e.} = N_{e1} + N_{e0}$	$p_e = N_{e1}/N_{e.}$	$RR = p_e/p_u$
$X = u$	N_{u1}	N_{u0}	$N_{u.} = N_{u1} + N_{u0}$	$p_u = N_{u1}/N_{u.}$	
Expanded	$Y^* = 1$	$Y^* = 0$		**Odds**	**Crude OR**
$X = e$	N_{e1}	$N_{e.}$		$odds_e^* = N_{e1}/N_{e.}$	$OR^* = p_e/p_u$
$X = u$	N_{u1}	$N_{u.}$		$odds_u^* = N_{u1}/N_{u.}$	

Mantel-Haenszel (M-H) OR from expanded cohort

- When there is a categorical confounder, Z, the Mantel-Haenszel adjusted RR is:

$$RR = \frac{\sum_k w^k RR^k}{\sum_k w^k}, \text{ where } w^k = \frac{N_{u1}^k N_{e.}^k}{N^k}.$$

- The Mantel-Haenszel OR of the *expanded* cohort:

$$OR^* = \frac{\sum_k w^{*k} OR^{*k}}{\sum_k w^{*k}}, \text{ where } w^{*k} = \frac{N_{u1}^k N_{e.}^k}{N^k + N_{.1}^k}.$$

- As we saw on previous slide, $OR^{*k} = RR^k$. Although weights are different, the weighted averages above are very close (we will show they estimate the same parameter).

Regression model for expanded cohort

Assume the relative risk (log-binomial) model for the probability of being a case:

$$\ln Pr(Y = 1 \mid X, Z) = \alpha + \beta X + \gamma Z.$$

(i.e., the adjusted RR for X is $\exp^{\beta}$)

Original	Expected $Y = 1$	Expected $Y = 0$	
$X = e$	$N_{e.}\exp\{\alpha + \beta + \gamma Z\}$	$N_{e.}(1 - \exp\{\alpha + \beta + \gamma Z\})$	
$X = u$	$N_{u.}\exp\{\alpha + \gamma Z\}$	$N_{u.}(1 - \exp\{\alpha + \gamma Z\})$	

Regression model for expanded cohort

Original	Expected $Y = 1$	Expected $Y = 0$	
$X = e$	$N_{e.}\exp\{\alpha + \beta + \gamma Z\}$	$N_{e.}(1 - \exp\{\alpha + \beta + \gamma Z\})$	
$X = u$	$N_{u.}\exp\{\alpha + \gamma Z\}$	$N_{u.}(1 - \exp\{\alpha + \gamma Z\})$	
Expanded	**Expected $Y^* = 1$**	**Expected $Y^* = 0$**	**Odds**
$X = e$	$N_{e.}\exp\{\alpha + \beta + \gamma Z\}$	$N_{e.}$	$\exp\{\alpha + \beta + \gamma Z\}$
$X = u$	$N_{u.}\exp\{\alpha + \gamma Z\}$	$N_{u.}$	$\exp\{\alpha + \gamma Z\}$

Logistic regression model of *expanded* data:

$$\ln \frac{Pr(Y^* = 1 \mid X, Z)}{1 - Pr(Y^* = 1 \mid X, Z)} = \alpha + \beta X + \gamma Z.$$

Adjusted OR from the *expanded data logistic regression model* ($\exp^{\beta}$) *is* adjusted RR.

Expanded data logistic regression: robust SE

The naïve SE from the expanded data logistic regression is too large:

(we have introduced noise by expanding the cohort).

A robust sandwich SE was proposed[1] to correct for this overestimation:

- "bread" is the naïve covariance matrix.
- "meat" is computed from the design matrix and residuals of the logistic regression.

1. Schouten et al. Risk Ratio and Rate Ratio Estimation in Case-Cohort Designs: Hypertension and Cardiovascular Mortality. *Stat. Med.* 1993;12(18):1733–1745.

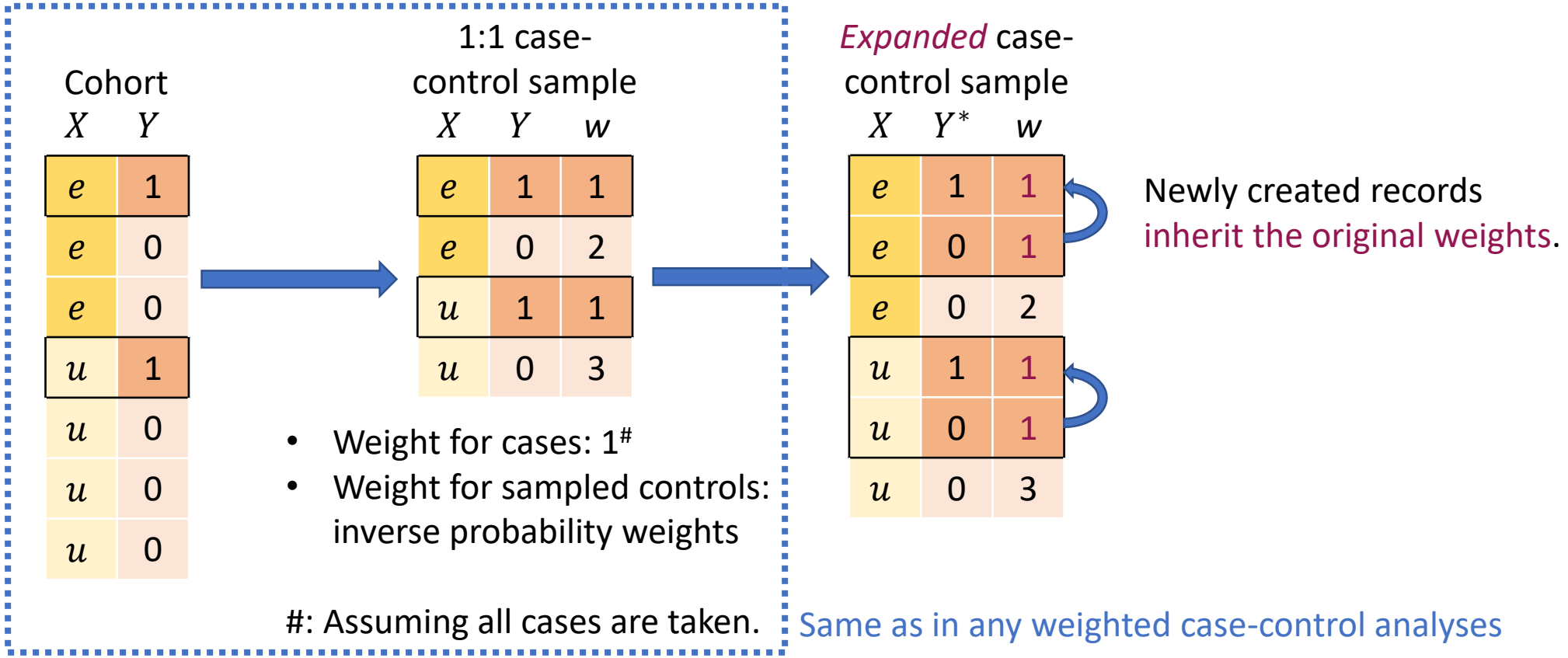
Doubling of cases in case-control studies

Doubling of cases in case-control studies

- A case-control sample can be regarded as "intentionally missing" data.
- If the sampling fractions are known, sampled controls can be up-weighted to "reconstruct" the cohort.
- The adjusted RR can now be estimated using an expanded data *weighted* logistic regression model.

- We derive a sandwich SE to adjust for the overestimation of variability:
  - "bread" is the naïve covariance matrix from the weighted logistic regression.
  - "meat" is computed from the design matrix and residuals of the weighted logistic regression.

How to assign weights?



R package `DoublingOfCases`

- Function `logit_db()` implements doubling of cases approach for cohort, cross-sectional and case-control design.

Cohort/cross-sectional data

X	Y
e	1
e	0
e	0
u	1
u	0
u	0
u	0

Expanded data logistic regression:

```
logit_db(y ~ x, data=dat)
```

Apply logistic regression to expanded cohort or cross-sectional data.

Case-control data

X	Y	w
e	1	1
e	0	2
u	1	1
u	0	3

Expanded data weighted logistic regression:

```
logit_db(y ~ x, data=dat_cc,
         weight_name="w")
```

Apply weighted logistic regression to expanded case-control data.

The doubling of cases step is handled within the function.

Applications

Simulation study

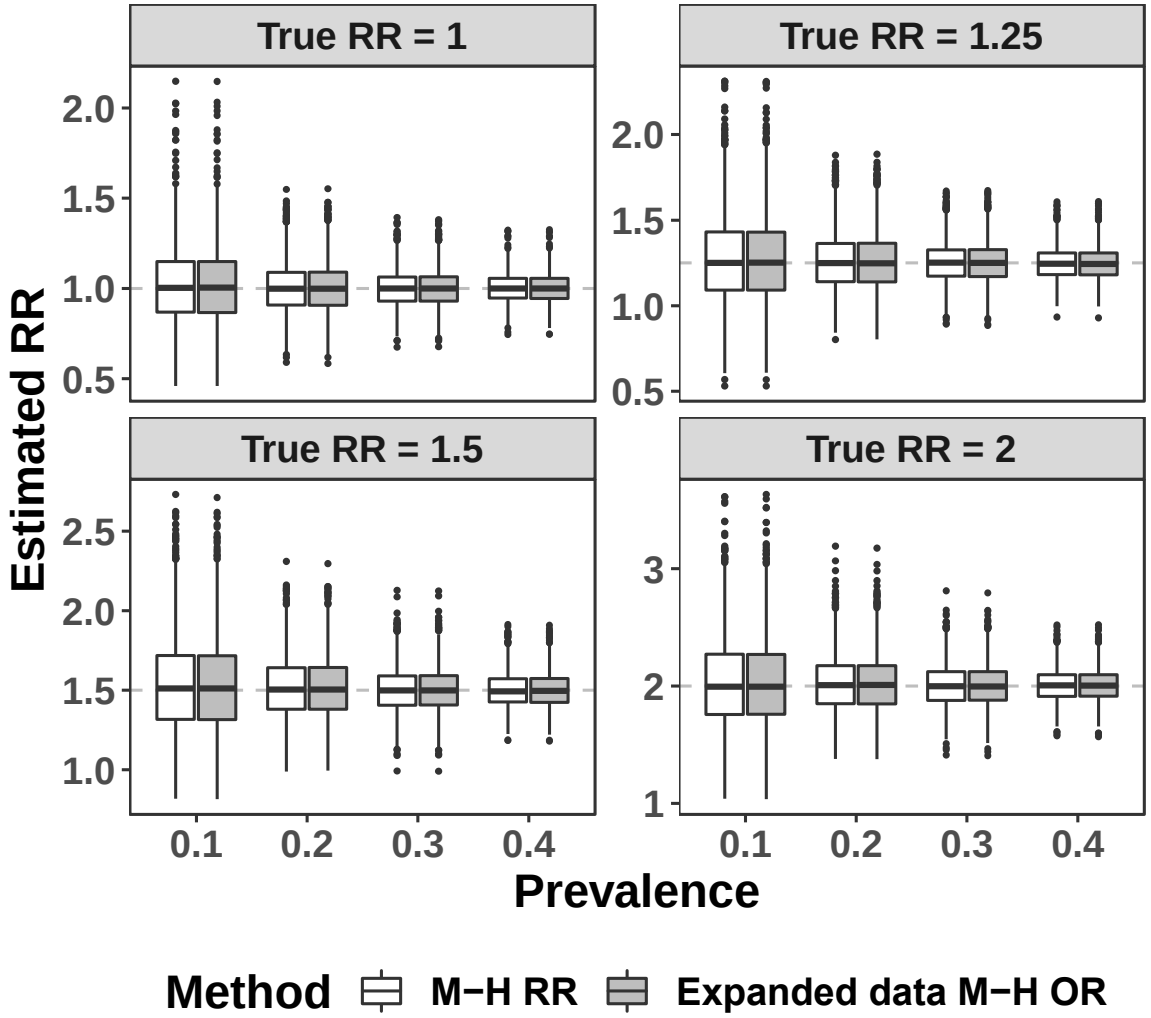
We generated data for a cohort of $N = 1000$ subjects with a binary outcome, binary exposure and binary confounder.

- Prevalence of disease ranging from 10% to 40%.
- True exposure effects: $RR = 1, 1.25, 1.5, 2.$

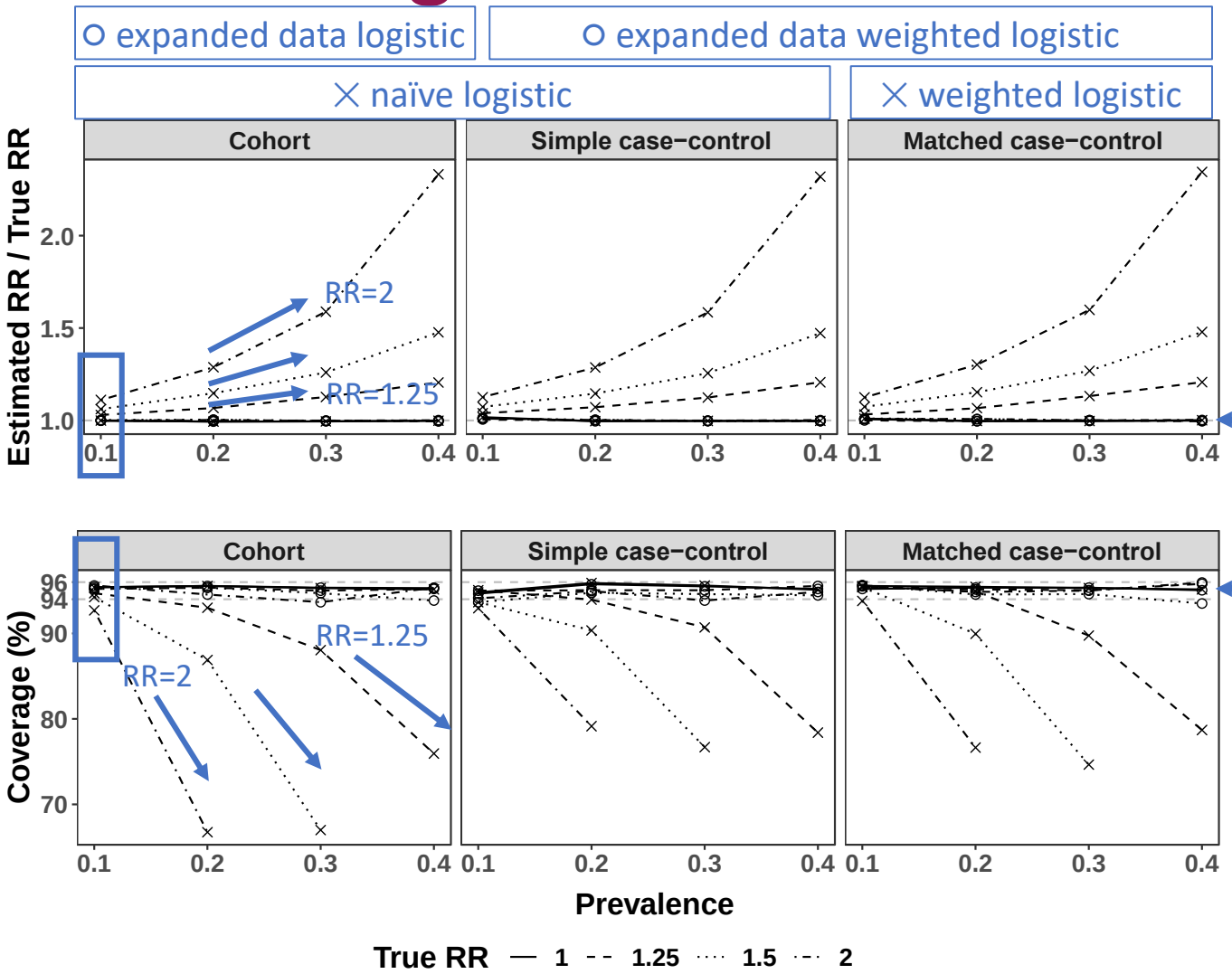
Sampled two 1:1 case-control samples from each cohort:

- one without matching (simple case-control)
- the other matched on the confounder (matched case-control).

Expanded data M-H OR is similar to M-H RR



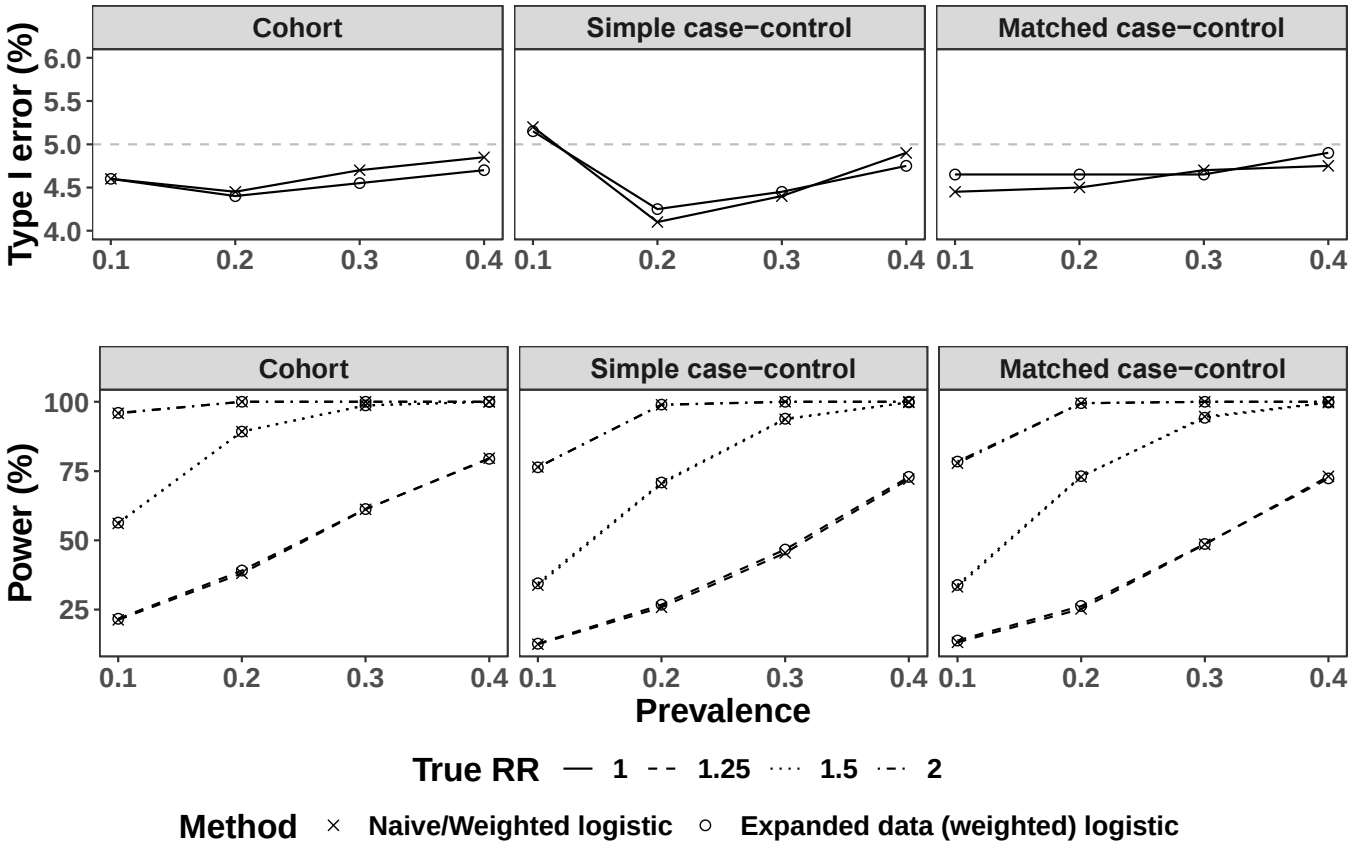
Expanded data logistic works well in estimating RR



Expanded data logistic (○): estimates are similar to true values

Expanded data logistic (○): coverage close to 95%

Both methods works well in detecting an effect



○ for expanded data logistic and
× for naïve logistic overlap

Illustrative example

We analysed preterm birth and other risk factors for neonatal jaundice in 547,466 singleton live births to Swedish women between 1992 and 2002,[2] where the mothers:

- were not alloimmunised and had no history of transfusion
- had complete information on the sex and prematurity of the infant, maternal age, BMI, parity and smoking status.

The outcome is rare: 21,441 (3.9%) infants had neonatal jaundice.

Crude OR associated with preterm was 28.0 but the crude RR was only 16.6.

Stronger association among multiparous mothers (crude OR=32.2 and crude RR=20.4) compared to nulliparous mothers (crude OR=23.4 and crude RR=13.1).

2. Lee et al. Haemolytic and nonhaemolytic neonatal jaundice have different risk factor profiles. *Acta Paediatr*. 2016;105(12):1444–1450.

OR overestimated RR despite rate event

- Estimated effect of risk factors from the full cohort:

Variables	Naive logistic	Log-binomial	Expanded data logistic
Preterm: nulliparous	23.5 (22.4, 24.5)	12.9 (12.5, 13.3)	13.0 (12.6, 13.4)
Preterm: multiparous	32.5 (30.8, 34.2)	20.1 (19.4, 20.9)	20.4 (19.6, 21.2)
Overweight: BMI ≥ 25	1.30 (1.26, 1.34)	1.20 (1.17, 1.23)	1.26 (1.23, 1.30)
Multiparous	0.50 (0.48, 0.52)	0.51 (0.50, 0.53)	0.51 (0.49, 0.53)

Overestimates RR

Estimates from these two methods are similar

Similar estimates from case-control sample

Similar findings from 1:2 case-control samples (64,323 births), matched on infant sex and maternal age.

Variables	Weighted logistic	Expanded data weighted logistic	Log-binomial estimates from full cohort
Preterm: nulliparous	23.8 (22.7, 24.9)	13.1 (12.3, 13.9)	12.9 (12.5, 13.3)
Preterm: multiparous	32.5 (30.9, 34.3)	20.5 (19.1, 21.9)	20.1 (19.4, 20.9)
Overweight: BMI ≥ 25	1.32 (1.26, 1.39)	1.28 (1.23, 1.33)	1.20 (1.17, 1.23)
Multiparous	0.50 (0.48, 0.52)	0.51 (0.49, 0.53)	0.51 (0.50, 0.53)

Summary

- The doubling-of-cases approach is simple, intuitive, and utilises the familiar logistic regression to estimate the adjusted RR.
- The doubling-of-cases approach applies to cohort, cross-sectional and case-control designs (by incorporating sampling weights).
- Researchers analysing binary outcomes should not feel restricted to report an OR.
- When a researcher chooses to report an OR, it is advisable to compare it with the RR to avoid exaggeration of effect sizes.