

Statistical Methods in Cancer Epidemiology using R

Janne Pitkaniemi

Faculty of Social Sciences, University of Tampere
Finnish Cancer Registry

Lecture 2b

janne.pitkaniemi@cancer.fi

Feb,17 2020

Confounding and effect modification

Consider another **factor** Z which is

- ▶ also a risk factor of the outcome,
- ▶ possibly associated with exposure X in study population,
- ▶ not a causal consequence of X .

⇒ Adjustment for possible *confounding* and evaluation of *effect modification* needed.

Example: OS vs. PN (cont'd)

Failure of treatment depends also on initial condition of patient, like extent and severity of disease.

Results stratified by initial diameter size of the stone:

Size	Trt	Fails	Npats	Fail-%	RD(%)
< 2 cm	OS	6	87	6.9	-6.1
	PN	35	270	13.0	
≥ 2 cm	OS	71	263	27.0	-4.3
	PN	25	80	31.3	

OS seems more successful in both subgroups, even though overall PN appeared better.

IS THERE A PARADOX?

Confounding

Solution: Treatment groups are not *comparable* w.r.t. initial size. Size of the stone is a *confounder* of the association between operation type and failure risk, because it is

- ▶ an *independent determinant* of outcome (failure), based on external knowledge,
- ▶ *statistically associated* with operation type in the study population,
- ▶ *not causally affected* by operation type.

This is an instance of “confounding by indication”:

- ▶ patient status affects choice of treatment → *bias* in comparing treatments.

This bias would be best to avoid in planning:

→ *randomized allocation* of treatments!

Stratified analysis

Stratification of cohort data with proportions

- ▶ at each level k of factor Z results are summarized:

Exposure to risk factor	Number of cases	Number of non-cases	Group size
yes	D_{1k}	C_{1k}	N_{1k}
no	D_{0k}	C_{0k}	N_{0k}
Total	D_{+k}	C_{+k}	N_{+k}

Stratum-specific incidence proportions by exposure group:

$$R_{1k} = \frac{D_{1k}}{N_{1k}}, \quad R_{0k} = \frac{D_{0k}}{N_{0k}}$$

Adjusted estimation of risk difference

- ▶ Let π_{jk} be true risk in exposure group j ($j = 0, 1$) as to X and stratum k ($k = 0, \dots, K$) of Z . Let also

$$\theta_k = \pi_{1k} - \pi_{0k}$$

be the risk difference in stratum k .

- ▶ Many approaches to combine stratum specific results into one summary estimator that adjusts for confounding.

These are all somehow weighted averages of stratum-specific estimators.

- ▶ Different weighting principles:
 - ▶ Maximum likelihood (ML),
 - ▶ Mantel-Haenszel (MH) weights,
 - ▶ Standardization either by external standard population or by “indirect” standardization.

Model-based adjustment of risk difference

Define generalized linear model for binary outcome with

- ▶ one binary exposure variable X and
- ▶ one binary stratifying factor or covariate Z (easily generalized to polytomous factors).

Random part: Number of cases D_{jk} in exposure group j ($j = 0, 1$) of X and level k ($k = 0, 1$) of Z is assumed to be binomially distributed

$$D_{jk} \sim \text{Binomial}(N_{jk}; \pi_{jk}),$$

Model-based adjustment of risk difference (cont'd)

Systematic part:

$$\pi_{jk} = \alpha + \beta X_j + \gamma Z_k,$$

where X_j is 0/1-indicator as before, and

- ▶ $Z_k = 1$ for level $k = 1$ of Z , otherwise $Z_k = 0$,
- ▶ $\alpha = \pi_{00} =$ baseline ("corner cell") risk,
- ▶ $\gamma = \pi_{01} - \pi_{00} = \pi_{11} - \pi_{10}$,
- ▶ $\beta = \pi_{10} - \pi_{00} = \theta_0 = \pi_{11} - \pi_{01} = \theta_1$,

How do we read this?

Model-based adjustment of risk difference (cont'd)

Implications of model definition

- ▶ the model assumes *homogeneity* of true risk difference θ associated with factor X (exposed vs. unexposed) across levels of Z : $\theta_1 = \theta_0 = \beta$,
- ▶ inclusion of Z in the model leads to adjustment of it when estimating the “true” effect θ of X ,
- ▶ γ = risk difference between levels 1 and 0 of Z ; this is same in both exposure groups ($j = 0, 1$)
 \Rightarrow homogeneity of the effect of Z is assumed, too.

Example. Treatment of renal calculi (cont'd)

- ▶ Define new variable
size = initial stone size (0 for small, 1 for large)
- ▶ Extended data matrix comprises four observational units (rows) and four variables (columns):

size	trt	fails	npats
0	1	6	87
0	0	35	270
1	1	71	263
1	0	25	80

- ▶ These may be read in as before, e.g.

```
library(Epi)
size <- c( 0, 0, 1, 1) ; trt <- c( 1, 0, 1, 0)
fails <- c( 6, 35, 71, 25)
npats <- c( 87, 270, 263, 80)
props <- fails/npats
```

Fitting model for adjusted risk difference

As before, but model formula supplemented by + size

```
RDmod2 <- glm( props ~ trt + size,  
               fam = binomial(link='identity'), w = npats)  
round( ci.lin(RDmod2), 3)
```

	Estimate	StdErr	z	P	2.5%	97.5%
(Intercept)	0.128	0.019	6.596	0.000	0.090	0.166
trt	-0.056	0.030	-1.888	0.059	-0.114	0.002
size	0.195	0.032	6.106	0.000	0.132	0.258

Reading the results:

- ▶ $\hat{\alpha} = \mathbf{0.128} = \hat{\pi}_{00}$, fitted baseline risk,
- ▶ $\hat{\gamma} = \mathbf{0.195}$, RD between large and small stones,
- ▶ $\hat{\beta} = -\mathbf{0.0561} [-\mathbf{0.114}, \mathbf{0.002}]$, estimated common treatment effect $\hat{\theta}$ for OS vs. PN.
= Weighted average of $\hat{\theta}_0 = -0.061$ and $\hat{\theta}_1 = -0.043$.

Effect modification

Homogeneity assumption – true differences were put equal:

$$\theta_k = \pi_{1k} - \pi_{0k} = \theta$$

across all levels k of covariate Z . *Is this realistic?*

Example. Is the true risk difference for treatment failure between OS and PN similar for small and big stones?

Empirical differences of failure proportions were

small stone: $\hat{\theta}_0 = 0.069 - 0.130 = \mathbf{-0.061}$

large stones: $\hat{\theta}_1 = 0.270 - 0.313 = \mathbf{-0.043}$

Is the contrast $-0.043 - (-0.061) = 0.018$ between these differences due to chance only, or is there essential *effect modification* present?

Modelling modification of risk difference

The random part is the same, but the systematic part is

$$\pi_{jk} = \alpha + \beta X_j + \gamma Z_k + \tau U_{jk}.$$

- ▶ $U_{jk} = X_j \times Z_k$, product of values of X and Z ,
- ▶ $\alpha = \pi_{00}$ = baseline ("corner cell") risk,
- ▶ $\beta = \pi_{10} - \pi_{00} = \theta_0$, $\gamma = \pi_{01} - \pi_{00}$,
- ▶ $\tau = \theta_1 - \theta_0 = (\pi_{11} - \pi_{01}) - (\pi_{10} - \pi_{00})$,
interaction parameter

τ describes, how much greater is the risk difference between levels 1 and 0 of risk factor X among those at level 1 of factor Z than in those at level 0.

Fitting model with modification

- ▶ Generation of an interaction of *product term*:
- ▶ $\text{trtsize} = \text{size} * \text{treat}$
- ▶ Expanded and rearranged data matrix:

fails	npats	size	trt	trtsize
35	270	0	0	0
6	87	0	1	0
25	80	1	0	0
71	263	1	1	1

```
size <- c( 0, 0, 1, 1) ;  
trt <- c( 0, 1, 0, 1);  
trtsize <- c(0,0,0,1)  
fails <- c( 35, 6, 25, 71)  
npats <- c( 270, 87, 80, 263)  
props <- fails/npats
```

Fitting model with modification (cont'd)

Fitting the model including the product term:

```
RDmod3 <- glm(props ~ trt + size + trtsize,  
              fam = binomial(link='identity'), w = npats)
```

Results and interpretation:

```
round( ci.lin(RDmod3)[ , -(3:4)] , 4)
```

	Estimate	StdErr	2.5%	97.5%
(Intercept)	0.1296	0.0204	0.0896	0.1697
trt	-0.0607	0.0340	-0.1273	0.0060
size	0.1829	0.0557	0.0737	0.2921
trtsize	0.0181	0.0678	-0.1147	0.1509

- ▶ $\hat{\beta} = -0.061 = \hat{\theta}_0$ = RD for OS vs. PN in small stones,
- ▶ $\hat{\gamma} = -0.183$ = RD btw large and small stones for OS.
- ▶ estimate [95 % CI] of the interaction parameter:

$$\hat{\tau} = \mathbf{0.0181[-0.115, 0.151]}$$

Final comments

- ▶ When risk ratio ϕ or odds ratio ψ is the parameter of interest, adjustment for confounding and evaluation of modification can be done by fitting an analogous binomial GLM with relevant link function.
- ▶ Modelling can easily be extended to cover one or more polytomous and/or continuous covariates. Flexible functional forms may be specified to describe the effects of the latter type of variables.
- ▶ Binomial models are not limited to grouped data but may be fitted on individual data with binary outcomes, too.
- ▶ With more complicated models, especially involving continuous variables, the identity link (sometimes log link, too) violates the basic range restriction: \ outcome probabilities π must remain within 0 and 1.