

# Statistical Analysis of Binary Composite Endpoints with Missing Data in Components

Daowen Zhang

*NC State University*  
DEPARTMENT OF STATISTICS



zhang@stat.ncsu.edu

<http://www4.ncsu.edu/~dzhang>

Joint work with Hui Quan, Ji Zhang & Laure Devlamynck  
Sanofi-Aventis

## Outline

1. Motivation
2. The Case of Two Components
3. The Case of Three Components
4. A Simulation Study
5. Application
6. Discussion

## 1. Motivation

A clinical trial:

Multicenter, randomized, double blind, parallel group trial for comparing an experimental treatment, Fondaparinux, to Enoxaparin on the prevention of venous Thromboembolism in patients with knee surgery

Primary endpoint – a composite endpoint consisting of

- Fatal or non-fatal PE (pulmonary embolism)
- DVT (deep vein thrombosis) through venograph between Day 5 and Day 11 – 30% missing

$$V = \begin{cases} 1 & \text{if PE or DVT} \\ 0 & \text{otherwise} \end{cases}$$

We would like to estimate, compare  $P[V = 1]$  for/between Fondaparinux and Enoxaparin.

## 2. The Case of Two Components

General data structure:

Patient type	$X$	$Y$	$V$
1	1	1	1
2	1	0	1
3	0	1	1
4	0	0	0
5	1	.	1
6	0	.	.
7	.	1	1
8	.	0	.
9	.	.	.

## Naive approaches:

1. **Naive 1:** Delete missing  $V$  (those  $V$  that cannot be determined, *i.e.* patient types 6, 8, 9)
  - Usually biased even under MCAR for  $X$  and  $Y$ .
  - Since when  $X = 0$  or  $Y = 0$ , the data are more likely to be deleted, the event rate  $P[V = 1]$  will be over-estimated.
  
2. **Naive 2:** Set  $V = 0$  for patient-types 6 and 8  $\Rightarrow$  the event rate  $P[V = 1]$  will be under-estimated.

**Correct approach: Maximum Likelihood**

$$X = \begin{cases} 1 & \text{Component 1 is an event} \\ 0 & \text{otherwise} \end{cases}$$

$$Y = \begin{cases} 1 & \text{Component 2 is an event} \\ 0 & \text{otherwise} \end{cases}$$

$$V = \begin{cases} 1 & \text{if } X = 1 \text{ or } Y = 1 \\ 0 & \text{otherwise} \end{cases}$$

Assume underlying structure for  $X$  and  $Y$ :

	$Y = 0$	$Y = 1$
$X = 0$	$\pi_{00}$	$\pi_{01}$
$X = 1$	$\pi_{10}$	$\pi_{11}$

Parameter of interest:

$$\gamma = P[V = 1] = 1 - P[X = 0, Y = 0] = 1 - \pi_{00}$$

$\Rightarrow$  Equivalent to estimating  $\pi_{00}$  for two treatment groups.

Let  $(X_i, Y_i)$  ( $i = 1, 2, \dots, n$ ) be the data and define

$$u_i^{jk} = I(X_i = j, Y_i = k), \text{ for } j, k = 0, 1$$

Then

$$u_i = (u_i^{00}, u_i^{01}, u_i^{10}, u_i^{11}) \sim \text{multinomial}\{1, \pi = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})\}.$$

The probability mass function when  $X_i, Y_i$  are fully observed is

$$P(u_i, \pi) = \pi_{00}^{u_i^{00}} \pi_{01}^{u_i^{01}} \pi_{10}^{u_i^{10}} \pi_{11}^{u_i^{11}}.$$

When  $X_i$ 's or  $Y_i$ 's are missing, define

$$n_{jk} = \sum_{i=1}^n I[X_i = j, Y_i = k]$$

$$n_{j.} = \sum_{i=1}^n I[X_i = j, Y_i = .]$$

$$n_{.k} = \sum_{i=1}^n I[X_i = ., Y_i = k]$$

Under MAR, the likelihood function of  $\pi = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$

$$\begin{aligned} L(\pi; Data) &= \pi_{00}^{n_{00}} \pi_{01}^{n_{01}} \pi_{10}^{n_{10}} \pi_{11}^{n_{11}} (\text{fully observed data}) \\ &\quad (\pi_{00} + \pi_{01})^{n_{0.}} (\pi_{10} + \pi_{11})^{n_{1.}} (X \text{ observed, } Y \text{ missing}) \\ &\quad (\pi_{00} + \pi_{10})^{n_{.0}} (\pi_{01} + \pi_{11})^{n_{.1}} (X \text{ missing, } Y \text{ observed}) \end{aligned}$$

- Obtain MLEs of  $\pi_{jk}$  by maximizing the log-likelihood

$$\ell(\pi; Data) = \log L(\pi; Data).$$

$\Rightarrow$  obtain MLE of  $\gamma = P[V = 1] = 1 - \pi_{00}$ . The variance of  $\hat{\gamma}_{ML}$  can be obtained using the observed information matrix from  $\ell(\pi; Data)$ .

- Asymptotically,

$$\hat{\gamma}_{ML} \sim N(\gamma, var(\hat{\gamma}_{ML})),$$

which can be used to make valid comparison between treatment groups.

- Direct maximization method such as Newton-Raphson for MLEs.
- No closed form solution in general.

Alternatively, EM algorithm for straightforward MLE calculation

**EM algorithm** for  $\hat{\pi}_{\text{ML}}$ : Given current estimate  $\pi^{(r)}$ , the  $Q$ -function is

$$\begin{aligned} Q(\pi|\pi^{(r)}) &= \text{E}\{\log P(u; \pi) | \text{Data}, \pi^{(r)}\} \\ &= n_{00}^{(r)} \log \pi_{00} + n_{01}^{(r)} \log \pi_{01} + n_{10}^{(r)} \log \pi_{10} + n_{11}^{(r)} \log \pi_{11}, \end{aligned}$$

where

$$\begin{aligned} n_{00}^{(r)} &= n_{00} + n_{0.} \left(1 - \delta_{0Y}^{(r)}\right) + n_{.0} \left(1 - \delta_{X0}^{(r)}\right), \\ n_{01}^{(r)} &= n_{01} + n_{0.} \delta_{0Y}^{(r)} + n_{.1} \left(1 - \delta_{X1}^{(r)}\right), \\ n_{10}^{(r)} &= n_{10} + n_{1.} \left(1 - \delta_{1Y}^{(r)}\right) + n_{.0} \delta_{X0}^{(r)}, \\ n_{11}^{(r)} &= n_{11} + n_{1.} \delta_{1Y}^{(r)} + n_{.1} \delta_{X1}^{(r)}, \end{aligned}$$

and

$$\begin{aligned} \delta_{jY}^{(r)} &= P[Y = 1 | X = j, \pi^{(r)}] = \pi_{j1}^{(r)} / (\pi_{j0}^{(r)} + \pi_{j1}^{(r)}), \\ \delta_{Xk}^{(r)} &= P[X = 1 | Y = k, \pi^{(r)}] = \pi_{1k}^{(r)} / (\pi_{0k}^{(r)} + \pi_{1k}^{(r)}). \end{aligned}$$

Maximizing  $Q(\pi|\pi^{(r)})$  w.r.t  $\pi \implies$  updates for  $\pi$ :

$$\pi_{00}^{(r+1)} = \frac{n_{00}^{(r)}}{n}$$

$$\pi_{01}^{(r+1)} = \frac{n_{01}^{(r)}}{n}$$

$$\pi_{10}^{(r+1)} = \frac{n_{10}^{(r)}}{n}$$

$$\pi_{11}^{(r+1)} = \frac{n_{11}^{(r)}}{n}$$

EM algorithm keeps iterating until convergence.

## Special Case: One component is completely observed

Let  $X$  be completely observed.

$$\begin{aligned}\gamma &= P[X = 1 \cup Y = 1] = P[X = 1] + P[X = 0 \cap Y = 1] \\ &= P[X = 1] + P[Y = 1|X = 0]P[X = 0]\end{aligned}$$

Denote

$$\begin{aligned}n_1 &= n(\# \text{ of patients}) & m_1 &= \sum_{i=1}^n I[X_i = 1] \\ n_2 &= \sum_{i=1}^n I[X_i = 0, Y_i \neq .] & m_2 &= \sum_{i=1}^n I[X_i = 0, Y_i = 1].\end{aligned}$$

The the MLE of  $\gamma$  is

$$\hat{\gamma} = \hat{p}_1 + \hat{p}_2(1 - \hat{p}_1) = \frac{m_1}{n_1} + \frac{m_2}{n_2} \frac{n_1 - m_1}{n_1}.$$

### 3. The Case of Three Components

Three binary components:  $X, Y, Z$

- 1 for event, 0 for non-event.
- $V = \max(X, Y, Z)$ .
- Parameter of interest  $\gamma = P[V = 1] = 1 - P[X = 0, Y = 0, Z = 0]$ .
- Each component can be 1, 0, or missing.
- A total of 26 ( $3^3 - 1$ ) possible patterns.

The value of  $V$  can be determined for some patients who had missing data in  $X$  or  $Y$  or  $Z$ . That is

$$V = \begin{cases} 1 & \text{if } X = 1 \text{ or } Y = 1 \text{ or } Z = 1 \\ 0 & \text{if } X = 0 \text{ and } Y = 0 \text{ and } Z = 0 \\ . & \text{otherwise} \end{cases}$$

As for the case of two components, direct analysis of  $V$  is difficult (often leads to bias):

- Cannot replace  $V = .$  by zero
- Cannot throw away  $V = .$
- Difficult to model  $P[V = .|Data]$

**Correct approach:** Analyzing  $X, Y, Z$  jointly using maximum likelihood approach!

Define for  $i, j, k = 0, 1$

$$\pi_{jkl} = P[X = j, Y = k, Z = l]$$

$$\pi_{jk.} = P[X = j, Y = k]$$

$$\pi_{j.l} = P[X = j, Z = l]$$

$$\pi_{.kl} = P[Y = k, Z = l]$$

$$\pi_{j..} = P[X = j]$$

$$\pi_{.k.} = P[Y = k]$$

$$\pi_{..l} = P[Z = l]$$

Similarly, define number of patients in each category. For example,

$$n_{jk.} = \sum_{i=1}^n I[X_i = j, Y_i = k, Z_l = .].$$

Under MAR, the likelihood function of  $\pi = (\pi_{jkl})_{j,k,l=0}^1$  is

$$L(\pi|Data) = \prod \pi_{jkl}^{n_{jkl}} \prod \pi_{jk.}^{n_{jk.}} \cdots \prod \pi_{j..}^{n_{j..}} \prod \pi_{.k.}^{n_{.k.}} \prod \pi_{..l}^{n_{..l}}$$

Direct maximization algorithms such as Newton-Raphson can be used to find MLEs of  $\pi_{jkl}$ 's, and hence the MLE of  $\gamma = 1 - \pi_{000}$ .

The asymptotic variance of  $\hat{\gamma} = 1 - \hat{\pi}_{000}$  can be derived from the observed information matrix.

When missing data pattern is monotonic:

Components

$X$	$Y$	$Z$
		?
	?	?

A closed form solution for the MLE's exists.

For example, if the missing data pattern is:

Components

$X$	$Y$	$Z$
	?	?

Then

$$\hat{\pi}_{j++} = \frac{n_{j++} + n_{j..}}{n}, \quad \frac{\hat{\pi}_{1kl}}{\hat{\pi}_{1++}} = \frac{n_{1kl}}{n_{1++}}, \quad \frac{\hat{\pi}_{0kl}}{\hat{\pi}_{0++}} = \frac{n_{0kl}}{n_{0++}}$$

Similar to two-component case, EM algorithm can be used to find MLEs for general missing data pattern.

## Between-Treatment Comparison

$\gamma_1, \gamma_2 =$  true event rates for Treatments 1 & 2.

MLEs:  $\hat{\gamma}_1, \hat{\gamma}_2$ , and their asymptotic variances  $\sigma_1^2/n_1, \sigma_2^2/n_2$ .

Between-treatment comparison can be based on

1. Log relative risk:

$$\log \frac{\hat{\gamma}_1}{\hat{\gamma}_2} \sim N \left\{ \log \frac{\gamma_1}{\gamma_2}, \frac{\sigma_1^2}{n_1 \gamma_1^2} + \frac{\sigma_2^2}{n_2 \gamma_2^2} \right\}.$$

2. Risk difference

$$\hat{\gamma}_1 - \hat{\gamma}_2 \sim N\{\gamma_1 - \gamma_2, \sigma_1^2/n_1 + \sigma_2^2/n_2\}.$$

## 4. A Simulation Study - Two Components

Denote

$$\pi_x = P[X = 1], \quad \pi_{y0} = P[Y = 1|X = 0].$$

The overall event rate for composite endpoint  $V = \max(X, Y)$  is

$$\gamma = \pi_x + \pi_{y0}(1 - \pi_x).$$

Assume MCAR for  $X$  and MAR for  $Y$ :

$$\pi_{mx} = P[X = .|X = j, Y = k] = \text{constant}$$

$$\pi_{my0} = P[Y = .|X = 0, Y = k] = \text{constant}$$

$$\pi_{my1} = P[Y = .|X = 1, Y = k] = \text{constant, different from } \pi_{my0}$$

$$P[X = ., Y = .] = 0.$$

*Simulation Results*

Pair	True $\gamma$	MLE	Naive 1	Naive 2
1	9.79	9.80	11.34	8.31
	9.79	9.77	13.08	7.60
2	14.50	14.49	16.72	12.36
	14.50	14.60	19.32	11.44
3	23.50	23.61	26.76	19.53
	23.50	23.31	28.14	18.89
4	28.00	28.04	31.01	22.60
	28.00	27.97	33.01	22.36

- In each pair of simulations, true  $\gamma$  is the same with different missing probabilities.
- Naive 1 = delete missing  $V$ ; naive 2 = set missing  $V$  to be zero.

*Simulation: Between-treatment Comparison*

Pair	True diff	MLE	Naive 1	Naive 2
1	1.00	1.01	1.00	0.92
	1.00	1.01	0.84	1.03
2	0.42	0.42	0.41	0.38
	0.42	0.42	0.36	0.42
3	1.19	1.19	1.15	1.11
	1.19	1.20	1.03	1.22

- In each pair of simulations,  $\gamma_2 - \gamma_1$  is the same with different missing probabilities.
- Naive 1 = delete missing  $V$ ; naive 2 = set missing  $V$  to be zero.

## 5. Application

A study for comparing Fondaparinux and Enoxaparin on prevention of thromboembolic complications in patients with total knee replacement (Bauer et al NEJM 2001).

- Treatment Duration: 5- 9 post-operative days
- Venograph: Day 5 to Day 11
- $X$ : fatal or non-fatal PE – fully observed
- $Y$ : DVT through venograph – 30% missing
- Primary endpoint  $V =$  composite endpoint of  $X$  and  $Y$ .

Sample size	Fondaparinux	Enoxaparin
$n$	517	517
$n_{00}$	316	262
$n_{01}$	44	97
$n_{10}$	0	1
$n_{11}$	1	1
$n_{0.}$	156	154
$n_{1.}$	0	2
$\hat{\gamma}_{ML}(\%)$	12.39	27.58
$\hat{\gamma}_{NA1}(\%)$	12.47	27.82
$\hat{\gamma}_{NA2}(\%)$	8.70	19.54
$\hat{\lambda}_{ML}(\%)$	0.449 (0.326, 0.619)	
$\hat{\lambda}_{NA1}(\%)$	0.448 (0.325, 0.617)	
$\hat{\lambda}_{NA2}(\%)$	0.446 (0.320, 0.619)	

Since  $X$  is fully observed and only  $Y$  has missing data, we have

$$\widehat{\gamma}_{na1} = \frac{n_{01} + n_{10} + n_{1.} + n_{11}}{n - n_0}$$

$$\widehat{\gamma}_{na2} = \frac{n_{01} + n_{10} + n_{1.} + n_{11}}{n}$$

$$\widehat{\gamma}_{ML} = \frac{n_{10} + n_{1.} + n_{11}}{n} + \frac{n_{01}}{n_{00} + n_{01}} \frac{n_{00} + n_{01} + n_0}{n}$$

$\implies$

$$\widehat{\gamma}_{na1} \geq \widehat{\gamma}_{ML} \geq \widehat{\gamma}_{na2}.$$

## Individual Components

After demonstrating treatment effect in composite endpoint  $V$ , we may want to compare treatment effect in individual composite endpoints  $X$  and  $Y$ :

- $X$  component:  $\pi_{10} + \pi_{11}$
- $Y$  component:  $\pi_{01} + \pi_{11}$
- Still need to model  $(X, Y)$  jointly.

Similarly for the case of three components.

## 5. Discussion

- We present a simple problem. It is so simple that it is easy to go wrong.
- Naive analyses are inconsistent.
- Jointly analyzing components provides consistent and efficient analysis under MAR; Agrees with ITT principle.
- It is possible to incorporate covariates.