

**STATISTICAL ISSUES IN CANCER
BIOMARKER ASSESSMENT.**

**Jeremy M G Taylor
Dept of Biostatistics
University of Michigan**

Biomarkers are a Big Deal in Cancer Research

NCI leadership has stated that biomarkers are the answer to speeding up progress on cancer research

Special issue of Clinical Cancer Research, June 2004.

Biomarkers, Surrogate End Points, and the Acceleration of Drug Development for Cancer Prevention and Treatment: An Update.

- Prologue
- 5 Review articles
- No statisticians were authors!

Article titles.

- Rationale for biomarkers and surrogate end points in mechanism-driven oncology drug development
- Reanalysis of approved cancer drugs: Old drugs, new tricks
- Colorectal adenomas: A prototype for the use of surrogate end points in the development of cancer prevention drugs
- CA-125 in clinical trial evaluation of new therapeutic drugs for ovarian cancer
- PDA-DT as a surrogate marker for evaluation of oncology drugs to treat prostate cancer

Where should statistics make a contribution?

1. Prognosis
2. Screening
3. Diagnosis
4. Stratification into trials
5. Selecting targetted therapy (Trial design)
6. Monitoring progression after treatment
7. Surrogate outcome in RCT
8. Auxiliary variable in RCT

Are statistical methods well developed?

1. Prognosis, Yes*; mainly - regression
2. Screening; Partially*
3. Diagnosis; Yes*, mainly - classification
4. Stratification into trials; Yes* - regression
5. Selecting targetted therapy; No
6. Monitoring progression after treatment; Partially* - longitudinal modelling
7. Surrogate outcome in RCT; Mainly*
8. Auxiliary variable in RCT; A bit*

* except for high dimensional data

Sources of data

- Immunohistochemistry
 - Flow cytometry
 - Gene expression
 - Mass Spec
 - etc
-
- subjectively assessed
 - automatically read
 - involve images
 - semi-automatic
 - involve non-open source software

Goals of Biomarker Studies

- Understanding the biology
- Utility in the clinic
- Lots in between
 - High throughput screening for interesting leads

Data structure

- Continuous
- Ordered categorical
- High dimensional
- Peaks
- Longitudinal
- Repeated measures

MULTIPLE BIOMARKERS

A common statement - A single biomarker will not be good enough to predict outcome, but combinations of biomarkers will be better at predicting outcome because each biomarker may represent a different aspect of the disease process.

- Is this wishful thinking because single biomarkers have not worked well enough?
- How does one combine multiple biomarkers?

Pancreatic Cancer Case-Control study - Two biomarkers with binary outcome.

- **CA19-9 and CA125 are biomarkers scored on samples from a case-control study as 0,1,2 (low, medium, high).**
 - **Cutpoints for CA19-9 were 11.2 and 28.6**
 - **Cutpoints for CA125 were 11.2 and 15**
- **Y measures a binary outcome (case/control status)**
- **Believe the probability of response to therapy should be a monotone function of CA19-9 and CA125**

Data

		CA125		
		Low	Med	High
Low		2/13	3/7	6/20
CA19-9	Med	1/10	1/3	8/10
High		12/16	10/13	47/49

Empirical Estimates

		CA125		
		Low	Med	High
Low		0.15 (1)	0.43 (3)	0.30 (4)
CA19-9	Med	0.10 (2)	0.33 (5)	0.80 (8)
High		0.75 (6)	0.77 (7)	0.96 (9)

Label

		Low	Med	High
Low		1	3	4
CA19-9	Med	2	5	8
High		6	7	9

Statistical Formulation

- Two ordered categorical covariates, (R, C)
- Binary outcome Y
- Goal - estimation of $p_{ij} = P(Y = 1 | R = i, C = j)$, utilizing the ordering

Notation

n_{ij} = number of observations in cell(i,j)

d_{ij} = number of responders in cell(i,j)

Monotone restriction $p_{ij-1} \leq p_{ij}$ and
 $p_{i-1j} \leq p_{ij}$

Statistical Question

- Does incorporating the partial ordering restriction give gains in efficiency?
- Does it induce bias?

Estimation methods

- Empirical $\hat{p}_{ij} = d_{ij}/n_{ij}$,
 $SE = (\hat{p}_{ij}(1 - \hat{p}_{ij})/n_{ij})^{1/2}$
- Modified Empirical followed by isotonic regression in 2 dimensions
- Gibbs-weak, Bayesian with independent priors

Prior(p) = $\prod_{ij} \text{Beta}(\alpha_{ij}, \beta_{ij})$

Gibbs sampling - draw p_{ij} from Beta distribution.

We use α_{ij}, β_{ij} such that prior means are ordered.

- Gibbs-or, Bayesian with strong ordered priors

Prior(p) $\propto \prod_{ij} \text{Beta}(\alpha_{ij}, \beta_{ij})$ I(order constraint satisfied)

Gibbs sampling - draw p_{ij} from truncated Beta distribution.

- Iso-Gibbs, Bayesian with isotonic regression (ad-hoc)

Prior(p) = $\prod_{ij} \text{Beta}(\alpha_{ij}, \beta_{ij})$

Hybrid Gibbs sampler, draw $\{p_{ij}\}$, isotonize, draw $\{p_{ij}\}$, isotonize, etc

Plots of estimated probabilities

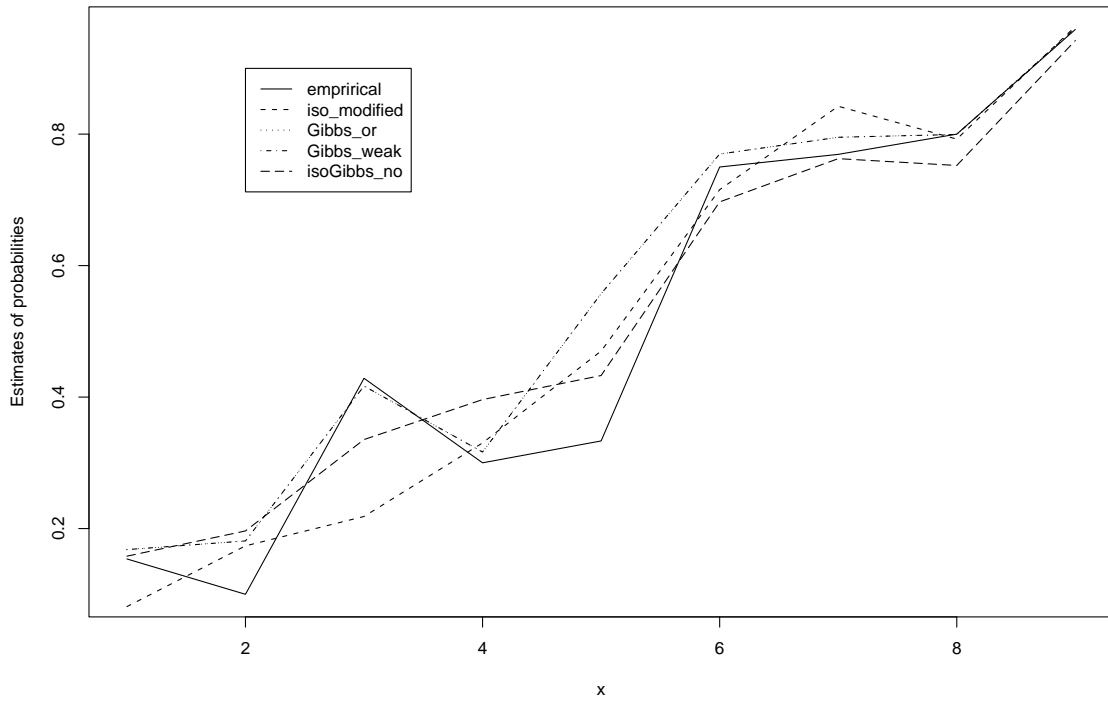


Figure 1: Estimated probabilities for 9 cells

Plots of estimated standard errors of the estimates of probabilities

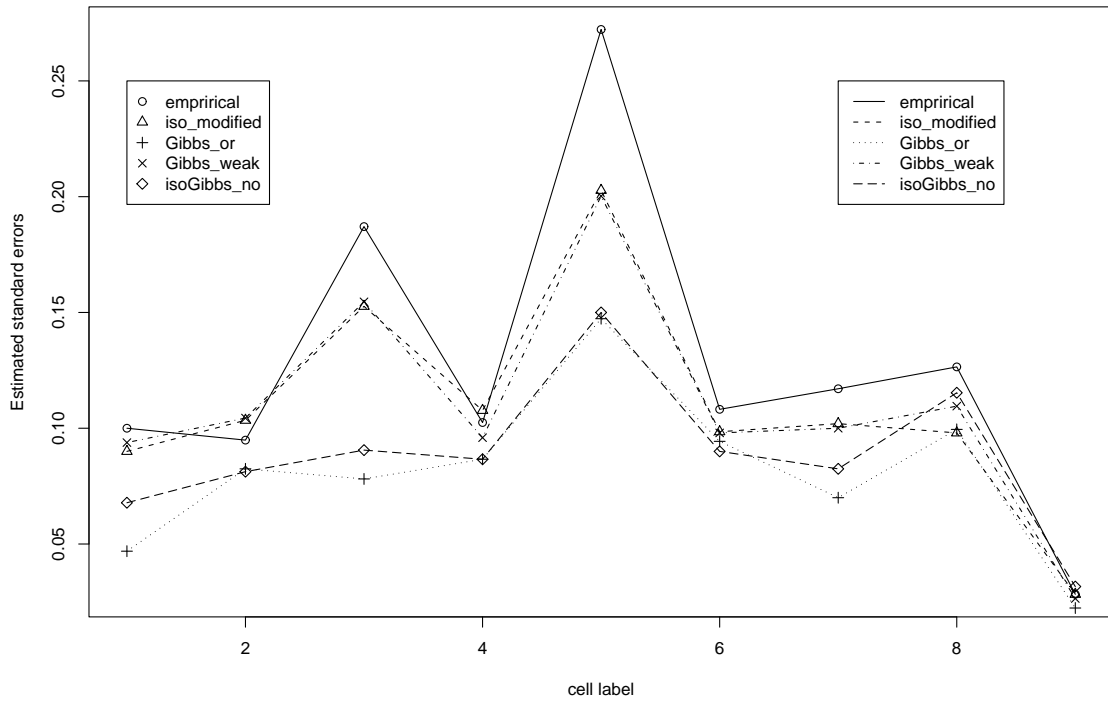


Figure 2: Estimated standard errors for 9 cells

Use estimates of p_{ij} to partition (R,C)

Iso-Gibbs estimates

	0	1	2
0	.15	.35	.40
1	.19	.43	.73
2	.70	.75	.94

Classification

	0	1	2
0	L	L	L
1	L	L	H
2	H	H	H

Simulation - 2 by 3 table

True values of p_{ij} :

0.4	0.5
0.6	0.9
0.8	0.91

Sample size = 300

Method	\overline{bias}	Effic	% Order	MSE
empirical	513	1	0.91	2149
Iso-modified	614	0.77	1	1730
Gibbs(weak)	497	0.92	0.91	1982
Gibbs(or)	1490	0.61	1	1521
IsoGibbs	877	0.69	1	1606

Method	\overline{cover}	Min cover
empirical	0.97	0.96
Iso-modified	0.94	0.93
Gibbs(weak)	0.94	0.89
Gibbs(or)	0.94	0.81
IsoGibbs	0.95	0.91

Conclusions

- Fully Bayesian methods with strong ordered priors gives large gains in efficiency, but can give bias.
- For fully Bayesian method estimates from adjacent cells tend to be pushed apart, but this is not the case for methods that utilize isotonic regression.
- Incorporating isotonic regression reduces bias, but still has gains in efficiency.
- The hybrid algorithm of adding isotonic step with Gibbs sampler seems to be quite effective.

Generalizations

- More than 2 ordered categorical biomarkers
- Missing biomarker data
- Censored survival time outcome
- Other covariates

SURROGATE AND AUXILIARY VARIABLE

- **Setting:** Clinical trial with a post-treatment or time-dependent biomarker
- **Biomarker as a surrogate endpoint -**
Replace true endpoint with surrogate endpoint
- **Biomarker as an auxiliary variable -**
Still base conclusions on true endpoint but use biomarker to help define distribution of true endpoint

Setting: Clinical trial

- Goal: estimation of marginal survival distribution of true endpoint
- Data
 - (t, δ) , censored event time
 - \mathbf{X} , time-independent covariates, e.g. treatment group
 - \mathbf{Y} , time-dependent biomarker
- View censored observations as missing data.
- Impute missing event times.
- From what distribution do you draw the imputes? (driven by data + model)

Example: Clinical trial in prostate cancer

- Longitudinal measurements of PSA
- PSA measured every 6 months post treatment
- True endpoint, Clinical recurrence
- Incorporate biomarker information into analysis of true recurrence times to recover information lost due to censoring.
- High PSA suggests clinical recurrence likely to occur soon.
- Low PSA suggests clinical recurrence not likely to occur soon.

- Incorporate biomarker information into analysis of failure times to recover information lost due to censoring.
- High PSA suggests clinical recurrence likely to occur soon.
- Low PSA suggests clinical recurrence not likely to occur soon.
- Potential benefits
 - Increase efficiency of survival distribution estimates
 - Increase power of test comparing 2 arms of trial
 - Reduce bias due to dependent censoring
 - * e.g., censoring related to auxiliary variables

MULTIPLE IMPUTATION

- Multiple imputation is a general technique for handling missing data
- Philosophy of Multiple Imputation. Once data is "filled-in", many different analyses can be tried.
- Imputation of the missing values from their predictive distribution (given the observed data) is repeated independently M times
⇒ M completed data sets
- Each completed data set is analyzed separately
- The M analyses results are combined using standard rules

- Combining analyses after multiple imputation
 - With M point and variance estimates,
 - * Final point estimate = average of M point estimates
 - * Final variance estimate =
Within-imputation +
Between-imputation variance estimates

Bayesian theoretical basis.

- Notation

- Q = Quantity of interest
- θ = parameters of the imputers model
- D_{obs} = observed data
- D_{mis} = missing data

- Two stages to the imputation

- 1. Draw parameters from their posterior distribution, $[\theta | D_{obs}]$
- 2. Draw missing value conditional on parameter and observed data, $[D_{mis} | \theta, D_{obs}]$

- Two stages to the analysis

- 1. Make inference based on $[Q | D_{mis}, D_{obs}]$
- 2. Repeat M times and combine results

No Covariate Case.

Non-parametric approach.

- Define Imputing Risk Set
 $R(j^+) = \{i : t_i > t_j, i = 1, \dots, n\}.$
- Kaplan-Meier Imputation (KMI) method
 - * Calculate K-M estimate of distribution of residual times from $R(j^+)$
 - * Draw an event time from this distribution.

- Result: KMI with large M reproduce the Kaplan-Meier estimator.
- Shares common ideas with the redistribute to the right algorithm.
- Provides a justification for multiple imputation of event times.

Extensions to incorporate covariates (auxiliary biomarker variables)

- Define Imputing Risk Set $R(j^+)$ as at-risk cases with similar relevant covariate values.
- Kaplan-Meier Imputation (KMI), draw from estimated residual distribution calculated from $R(j^+)$.
- Applies to both baseline and time-dependent biomarkers.

Single Baseline Covariate (X).

Define Imputing Risk Set (nearest neighbourhood) for each censored case as observations with similar values of X amongst those at risk.

Multiple Covariates (X_1, \dots, X_p).

Reduce to 2 dimensions, RS_f and RS_c .

- $RS_f = \Sigma \beta_j X_j$
= Risk score from failure time model
= linear combination most associated with event time in a Cox model.
- $RS_c = \Sigma \gamma_j X_j$
= Risk score from censoring time model = linear combination most associated with censoring time in a Cox model.

- Choose nearest neighbours to censored case j with small values of $\omega_f(RS_f(j) - RS_f(k))^2 + \omega_c(RS_c(j) - RS_c(k))^2$.
where $\omega_f + \omega_c = 1$.
- Nearest neighbours are in an ellipse centered at RS_f, RS_c for the censored case.

- KMI method
 - * Calculate estimated residual time distribution from set of neighbours
 - * Draw an event time from this distribution

- Why use both RS_f and RS_c ?
 - * Within each neighbourhood, censoring and failure time are independent, thus estimated residual time distribution is “good”, even if one of the two models is misspecified.
 - * Double robustness property

Issues

- how big should the neighbourhood be?
 - * 5, 10, 20, depends on follow-up.
- what value should be used for ω_f ?
 - * 0.5 - 0.8

Simulation Study Results

- Method works well.
- Can reduce bias due to dependent censoring.
- Can gain efficiency, compared to not using biomarker, but gains tend to be small.
- Appears to be as or more efficient than Inverse Probability of Censoring Weighted methods.

Parametric approach - Joint models for survival and longitudinal data

– Longitudinal model for biomarker

$$* Y_i(t_{ij}) = Z_i(t_{ij}) + e_{ij}$$

$$* Z_i(t) = X\beta + a_i + b_it$$

– Hazard model (proportional hazards)

$$* \lambda(t) = \lambda_0(t) \exp(\phi Z_i(t) + \gamma X_i)$$

- Estimation via MCMC or MLE.
- For censored subject draw T_j from residual time distribution S_R .
- $S_R(\tau) = P(T_j > \tau | T_j > t_j, X_j, Y_j)$.
- This distribution is determined by the joint model.
- Repeat draws of T_j from $S_R(\tau)$ multiple times
- Careful to limit the range of the imputes to avoid extrapolation.

- Methods of analysis for augmented data sets (i.e., primary analyses)
 - * Kaplan-Meier estimates for treatment and placebo groups
 - * Cox regression with treatment indicator as predictor

Discussion.

- Multiple-imputation approach has potential for
 - * increasing efficiency
 - * adjusting for dependent censoring

- Expect greater benefits with more accurate and/or more frequently measured biomarker variables

Auxiliary variables

- Main interest is in $[T]$, event time distribution.
- Specify joint distribution of $[T, Y]$ as $[Y][T|Y]$
- More natural than $[T, Y] = [T][Y|T]$
- Obtain $[T]$ from $[T, Y]$ by integrating out Y , via multiple imputation

Surrogate endpoint

- Base inference on $[Y|X]$ instead of $[T|X]$, where X denotes treatment group
- Key surrogacy property, $[T|Y, X] = [T|Y]$
- Understanding joint distribution of T and Y is necessary to assess surrogacy.
- Not sufficient to know $[Y|X]$ and $[T|Y]$
- Joint models can also lead to estimates of quantities such as the proportion of treatment effect explained and other measures of surrogacy, and can generalize to multiple trials