NONPARAMETRIC ADJUSTMENT FOR MEASUREMENT ERROR IN TIME TO EVENT DATA: APPLICATION TO RISK PREDICTION MODELS

Malka Gorfine
Tel Aviv University, Israel

Joint work with

Danielle Braun and Giovanni Parmigiani,
Department of Biostatistics, Harvard School of Public Health

Setting

Risk prediction model has first been developed based on **error-free time to event data**, and subsequently implemented in practical setting where **time to event data can be error-prone**.

Motivation I

Mendelian risk prediction models in genetic counseling:

- Calculating the probability that an individual carries a cancer causing inherited mutation based on his/her **family history**.
- Predicting the absolute risk of developing the disease over time given his/her mutation-carrier status and **family history**.

These models are in wide clinical use and web-based patient-oriented tools: breast cancer, ovarian cancer, Lynch syndrome, pancreatic cancer, melanoma etc.

These models were developed based on error-free data.

Studies about accuracy of self-reported family history show that sensitivity and specificity for reported disease status vary by degree of relative and cancer type.

65% of truly affected are reported affected

2% of truly unaffected are reported as affected

Breast cancer:

Disease status, sensitivity 65% - 95%; specificity 98% - 99%.

Age of diagnosis was misreported for 3.1% of relatives, average of 4.5 years between the true and misreported ages (Mai et al 2011; Ziogas and Anton-Culver, 2003).

Ovarian cancer:

Age of diagnosis was misreported for 4.2% of relatives, average of 4.2 years between the true and misreported ages (Ziogas and Anton-Culver, 2003).

Misreporting of family history, especially in disease status, leads to distortions in predictions (Katki, 2006).

Q: Is it possible to develop prediction models based on error-prone data?

A: Not in the context of Mendelian risk prediction models which relies on penetrance estimates from the literature, based on error-free data.

disease probability given carrier status

Motivation II

Survival prediction models:

Time to progression – length of time until the disease gets worse or spread

In some disease settings, such as cancer, TTP is one of the predictors for survival.

Assume a model has been developed based on **error-free TTP**.

In practice, **TTP is error-prone**:

- Tumor assessment is done using imaging, which varies by observers.
- Scans are taken at regularly intervals.

The current setting vs the common settings

The usual measurement error setting:

- ✓ Error-prone covariate observed in the main study.
- ✓ The goal is estimating the relationship between the outcome and the true covariate.

Current setting:

- ✓ The relationship between the outcome and the true covariate is known.
- ✓ The goal is to use this model for risk estimation based on an error-prone covariate.

Naïvely using the error-prone covariate will lead to biased results.

The data:

Y - outcome

T^o - the true failure time

C - the true right-censoring time

Error-free predictor: $H = (T, \delta)$ (for simplicity, one relative)

where $T = \min(T^o, C)$ $\delta = I(T^o \leq C)$

Example: $Y = 0$ or 1 and T^o the mother's age at onset.

Error-free predictor: $H = (T, \delta)$

Error-prone predictor: $H^* = (T^*, \delta^*)$

Example: the counselee doesn't know that his/her relative had the disease, or the correct age at onset.

Assumption: We have a validation study with

$$H = (T, \delta) \quad \text{and} \quad H^* = (T^*, \delta^*)$$

but no need for the outcome Y.

The risk prediction model: $\Pr(Y \mid H)$

Our goal is estimating $\Pr(Y \mid H^*)$

Our goal is estimating $\Pr\left(Y \mid H^*\right)$

Main idea:

$$\mathrm{P}\left(Y \mid H^*\right) = \int_H \mathrm{P}\left(Y, H \mid H^*\right) dH$$

$$= \int_H \mathrm{P}\left(Y \mid H, H^*\right) \mathrm{P}\left(H \mid H^*\right) dH$$

$$= \int_H \mathrm{P}\left(Y \mid H\right) \mathrm{P}\left(H \mid H^*\right) dH$$

Assumptions:

surrogacy assumption

- H^* contains no information on predicting Y beyond H.
- The measurement error model $\mathrm{P}\left(H \mid H^*\right)$ is transportable.

A non parametric estimator of $\Pr\left(H \mid H^*\right)$

The distribution is left unspecified

Assume, for simplicity, one family member → $H=\left(T,\delta\right) \qquad H^*=\left(T^*,\delta^*\right)$

Main idea:

$$\mathrm{P}\left(T,\delta \mid T^*,\delta^*\right)$$

$$=\lambda\left(T \mid T^*,\delta^*\right)^{\delta} S\left(T \mid T^*,\delta^*\right) h\left(T \mid T^*,\delta^*\right)^{1-\delta} G\left(T \mid T^*,\delta^*\right)$$

Conditional hazard and survival of true failure time

Conditional hazard and survival of true censoring time

Assumptions:

Conditional independence of event and censoring times given H^*.

$$\mathrm{P}\left(T,\delta \mid T^*,\delta^*\right)$$

$$= \lambda\left(T \mid T^*,\delta^*\right)^{\delta} S\left(T \mid T^*,\delta^*\right) h\left(T \mid T^*,\delta^*\right)^{1-\delta} G\left(T \mid T^*,\delta^*\right)$$

These hazards and survival functions can be estimated non-parametrically by using the validation data – a large study population that does not involve the counselee.

Validation data

$$T_i = \min\left(T_i^o, C_i\right) \quad \delta_i = I\left(T_i^o \le C_i\right) \quad i = 1,\ldots,n$$

dependent individuals

$$H_i = \left(T_i,\delta_i\right) \quad H_i^* = \left(T_i^*,\delta_i^*\right)$$

Use kernel smoothed Kaplan-Meier estimator (Beran, 1981).

Kernel smoothed Kaplan-Meier estimator (Beran, 1981):

Nadaraya-Watson weight

$$W_i\left(t;b_{nl},l\right)=I\left(\delta_i^*=l\right)K\left(\frac{t-T_i^*}{b_{nl}}\right) \qquad i=1,\ldots,n \quad l=0,1$$

Bandwidth sequences

Known kernel function

$$\hat{S}\left(t\,|\,t^*,l\right)=\prod_{T_i\le t,\delta_i=1,\delta_i^*=l}\left(1-\frac{W_i\left(t^*;b_{nl},l\right)}{\sum_{j=1,\delta_j^*=l}^{n}W_j\left(t^*;b_{nl},l\right)I\left(T_j\ge T_i\right)}\right)$$

and we get $\hat{S}\left(t\,|\,t^*,0\right)$, $\hat{S}\left(t\,|\,t^*,1\right)$ for $t,t^*\in(0,\tau]$.

For estimating the survival function of the censoring time, apply the above while treating the censoring times as events and the event times as censoring.

Summary

provided

$$\hat{\mathrm{P}}\left(Y \mid H^*\right) = \int_H \mathrm{P}\left(Y \mid H\right) \hat{\mathrm{P}}\left(H \mid H^*\right) dH$$

$$H = \left(H_1, \ldots, H_R\right)$$

$$H^* = \left(H_1^*, \ldots, H_R^*\right)$$

$$\mathrm{P}\left(H \mid H^*\right) = \prod_{j=1}^{R} \mathrm{P}\left(H_j \mid H_j^*\right)$$

$$\hat{\mathrm{P}}\left(H \mid H^*\right) = \prod_{j=1}^{R} \hat{\mathrm{P}}\left(H_j \mid H_j^*\right)$$

$$\hat{\mathrm{P}}\left(H_j \mid H_j^*\right) = \hat{\mathrm{P}}\left(T_j, \delta_j \mid T_j^*, \delta_j^*\right)$$

$$= \hat{\lambda}\left(T_j \mid T_j^*, \delta_j^*\right)^{\delta_j} \hat{S}\left(T_j \mid T_j^*, \delta_j^*\right) \hat{h}\left(T_j \mid T_j^*, \delta_j^*\right)^{1-\delta_j} \hat{G}\left(T_j \mid T_j^*, \delta_j^*\right)$$

provided

$$\hat{\mathrm{P}}\left(Y \mid H^{*}\right)=\int_{H} \mathrm{P}\left(Y \mid H\right)\hat{\mathrm{P}}\left(H \mid H^{*}\right)dH$$

In case integrating over all possible values of H is computational challenging, a Monte-Carlo estimator can be used, by sampling

$$H^{(1)},\ldots,H^{(B)}$$

From $\hat{\mathrm{P}}\left(H \mid H^{*}\right)$ and the final proposed estimator is given by

$$\hat{\mathrm{P}}\left(Y \mid H^{*}\right)=\frac{1}{B}\sum_{b=1}^{B}\hat{\mathrm{P}}\left(Y \mid H^{(b)}\right)$$

Application: Mendelian Risk Prediction Model

A counselee provides information on R relatives:

$$H_i^* = \left(T_i^*, \delta_i^*\right) \text{ instead of } H_i = \left(T_i, \delta_i\right) \qquad i = 1, \ldots, R$$

Let

$$\gamma_i = \left(\gamma_{i1}, \ldots, \gamma_{iM}\right), \quad \gamma_{im} = 0 \text{ or } 1$$

$\gamma_{im} = 1$ indicates carrying the genetic variant that confer disease risk

Aims:

- Estimating $\mathrm{P}\left(\gamma_0 \mid H_0, H_1^*, \ldots, H_R^*\right)$
- Estimating $\mathrm{P}\left(T_0^o > t \mid H_0, H_1^*, \ldots, H_R^*\right)$

Carrier probability $\mathrm{P}(Y \mid H) = \mathrm{P}(\gamma_0 \mid H_0, H_1, \ldots, H_R)$

Write

$$\mathrm{P}(\gamma_0 \mid H_0, H_1, \ldots, H_R) = \frac{\mathrm{P}(\gamma_0) \sum_{\gamma_1, \ldots, \gamma_R} \prod_{i=0}^{R} \mathrm{P}(H_i \mid \gamma_i) \mathrm{P}(\gamma_1, \ldots, \gamma_R \mid \gamma_0)}{\sum_{\gamma_0} \mathrm{P}(\gamma_0) \sum_{\gamma_1, \ldots, \gamma_R} \prod_{i=0}^{R} \mathrm{P}(H_i \mid \gamma_i) \mathrm{P}(\gamma_1, \ldots, \gamma_R \mid \gamma_0)}$$

conditional independence of family members' phenotype given their genotypes.

BRCAPRO estimate it via meta-analysis, and family history information is verified using medical records

and in practice $\mathrm{P}(Y \mid H^*) = \mathrm{P}(\gamma_0 \mid H_0, H_1^*, \ldots, H_R^*)$

is naively being used. We propose

$$\hat{\mathrm{P}}(\gamma_0 \mid H^*) = \int_H \mathrm{P}(\gamma_0 \mid H) \hat{\mathrm{P}}(H \mid H^*) dH$$

Survival probability $\mathrm{P}\left(Y \mid H\right) = \mathrm{P}\left(T_0^o > t \mid H_0, H_1, \ldots, H_R, \gamma_0\right)$

Write

$$\mathrm{P}\left(T_0^o > t \mid H_0, H_1, \ldots, H_R, \gamma_0\right)$$

$$= \frac{\mathrm{P}\left(T_0^o > t \mid \gamma_0\right) \sum_{\gamma_1,\ldots,\gamma_R} \prod_{i=1}^{R} \mathrm{P}\left(H_i \mid \gamma_i\right) \mathrm{P}\left(\gamma_1, \ldots, \gamma_R \mid \gamma_0\right)}{\sum_{\gamma_1,\ldots,\gamma_R} \prod_{i=0}^{R} \mathrm{P}\left(H_i \mid \gamma_i\right) \mathrm{P}\left(\gamma_1, \ldots, \gamma_R \mid \gamma_0\right)}$$

and in practice $\mathrm{P}\left(Y \mid H^*\right) = \mathrm{P}\left(\gamma_0 \mid H_0, H_1^*, \ldots, H_R^*\right)$

is naively being used. We propose

$$\hat{\mathrm{P}}\left(T_0^o > t \mid H^*\right) = \int_H \mathrm{P}\left(T_0^o > t \mid H\right) \hat{\mathrm{P}}\left(H \mid H^*\right) dH$$

Simulation Study

- Setting: $\mathrm{P}\left(Y \mid H\right) = \mathrm{P}\left(\gamma_0 \mid H_0, H_1, \ldots, H_R\right)$ with single gene BRCA1
- Two datasets were generated, one to model the measurement error distribution, and the other represents the counselees.
- The BRCA1 carrier probability 0.006098.
- The penetrance function $\mathrm{P}\left(H \mid \gamma\right)$ from BRCAPRO version 2.08.
- Normal censoring, mean 55, SD 10.
- **Measurement error in disease status**: sen=0.954, spec=0.974; sen=0.649 and spec=0.990.
- **Measurement error in age**:

$$T^* = T + \varepsilon \quad \varepsilon \sim N\left(0, \sigma^2\right) \quad \sigma = 1,3,5$$

$$T^* = TU \quad U \sim Exp\left(1\right)$$

50,000 counselees

100,000 families, each with 5 members (mother, father, 3 daughters)

Results

Table 1: Mendelian Risk Prediction Simulation Results. MSEP and O/E improve using the adjusted proposed method, ROC-AUC either improves or remains the same depending on the setting.

Counselee	Sens/Spec	Error in Age	$\sqrt{MSEP^{\dagger}} * 1000$			O/E			ROC-AUC		
			Error-Free	Error-Prone	Adjusted	Error-Free	Error-Prone	Adjusted	Error-Free	Error-Prone	Adjusted
Mother	0.954, 0.974	a: $N(0,5^2)$	0.0000	19.1351	16.7405	0.9773	0.8190	0.9712	0.8160	0.8090	0.8086
		a: $N(0,3^2)$	0.0000	18.1006	15.9490	0.9773	0.8280	0.9746	0.8160	0.8098	0.8078
		a: $N(0,1^2)$	0.0000	17.5526	15.6430	0.9773	0.8327	0.9746	0.8160	0.8102	0.8115
		m: $exp(1)$	0.0000	43.2855	21.3037	0.9833	0.6063	0.9484	0.8145	0.7185	0.8020
	0.649, 0.990	a: $N(0,5^2)$	0.0000	21.3122	20.8859	0.9773	1.0466	0.9783	0.8160	0.7814	0.7803
		a: $N(0,3^2)$	0.0000	20.7947	20.5099	0.9773	1.0556	0.9792	0.8160	0.7821	0.7815
		a: $N(0,1^2)$	0.0000	20.5213	20.1459	0.9773	1.0604	0.9737	0.8160	0.7826	0.7818
		m: $exp(1)$	0.0000	36.0314	23.8184	0.9817	0.8026	0.9565	0.8155	0.7140	0.7752
Daughter	0.954, 0.974	a: $N(0,5^2)$	0.0000	18.8437	16.5614	0.9719	0.8166	0.9680	0.8171	0.8070	0.8082
		a: $N(0,3^2)$	0.0000	17.7033	15.6918	0.9719	0.8256	0.9659	0.8171	0.8083	0.8086
		a: $N(0,1^2)$	0.0000	17.1421	15.1775	0.9719	0.8301	0.9680	0.8171	0.8093	0.8083
		m: $exp(1)$	0.0000	43.4717	21.0365	0.9785	0.6028	0.9445	0.8162	0.7146	0.7976
	0.649, 0.990	a: $N(0,5^2)$	0.0000	20.1613	20.0763	0.9719	1.0573	0.9872	0.8171	0.7895	0.7862
		a: $N(0,3^2)$	0.0000	19.6759	19.5498	0.9719	1.0661	0.9834	0.8171	0.7913	0.7904
		a: $N(0,1^2)$	0.0000	19.4507	19.1871	0.9719	1.0708	0.9803	0.8171	0.7928	0.7911
		m: $exp(1)$	0.0000	35.2381	23.0851	0.9760	0.8087	0.9535	0.8165	0.7229	0.7745

† MSEP: difference between (adjusted) error-prone and error-free predictions.

a: indicates a classical additive model; $T^* = T + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$, m: indicates a multiplicative measurement error model, $T^* = TU$, $U \sim exp(\lambda)$.

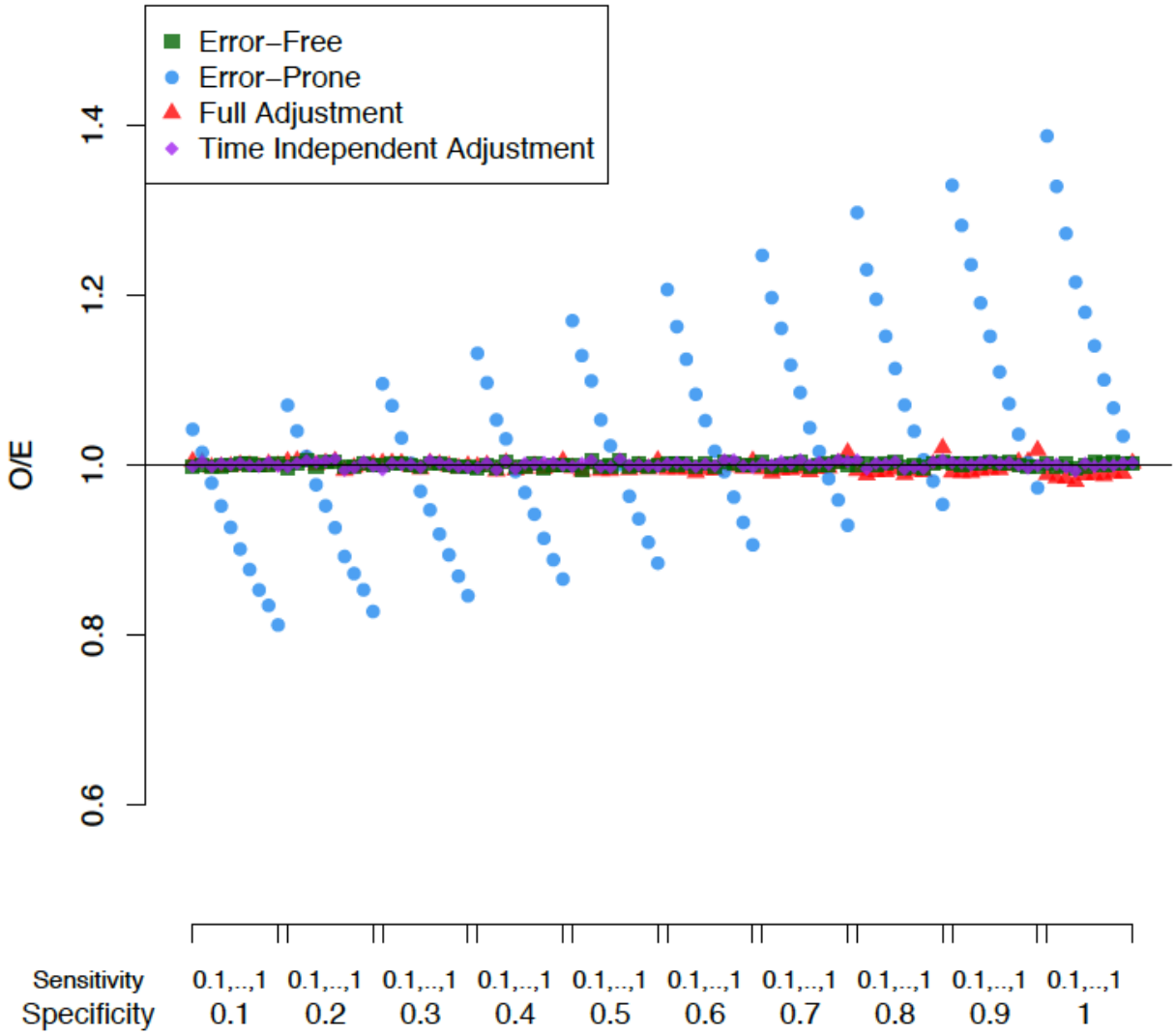
$$\text{O/E} = \sum_i I\left(\gamma_{0i} = 1\right) / \sum_i \hat{P}_i$$

$$\text{MSEP} = n^{-1} \sum_i \left\{ \hat{P}_i - \hat{\text{P}}_i\left(\gamma_0 \mid H\right) \right\}^2$$

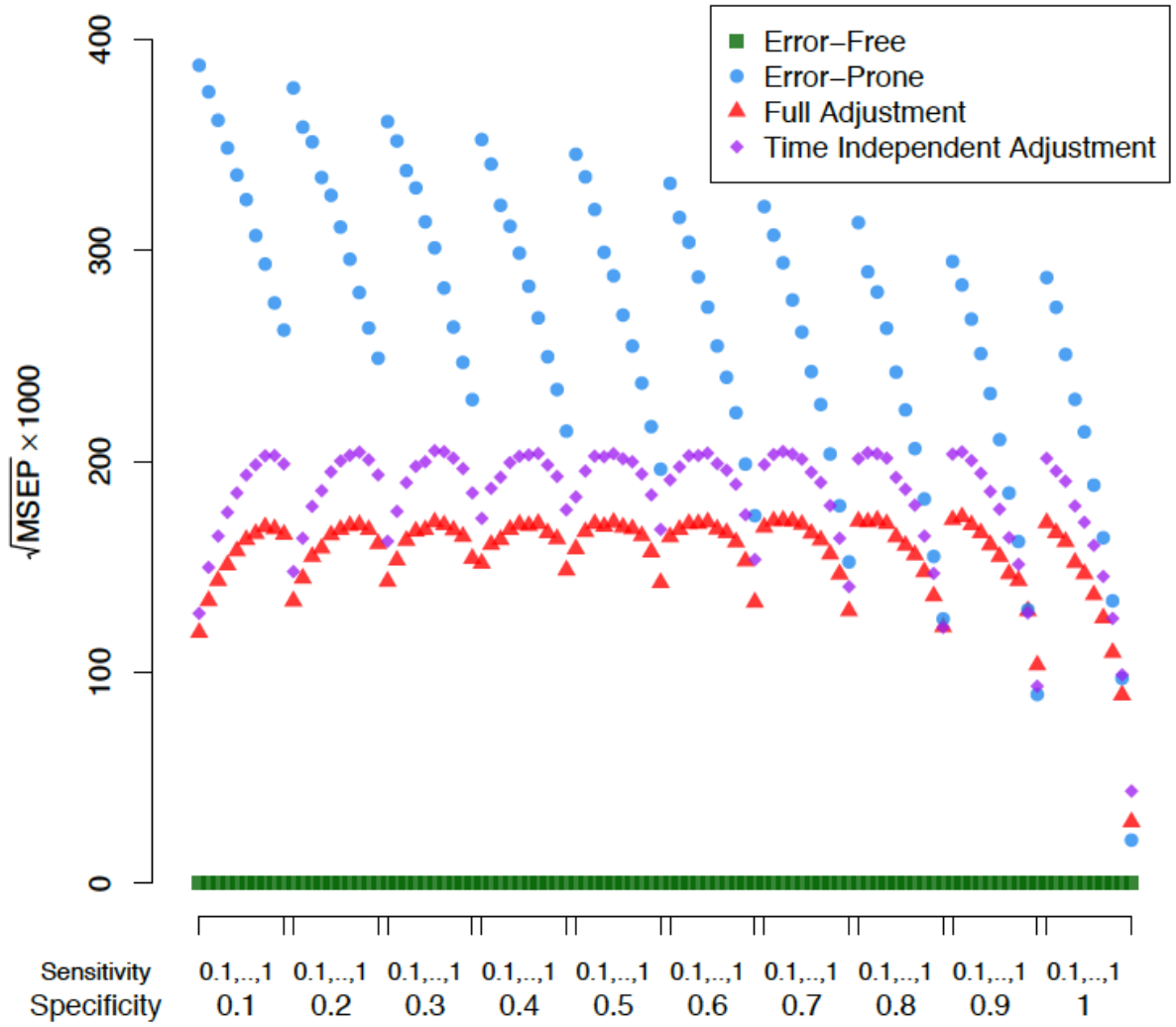
Summary of simulation results

- We are able to eliminate almost all the bias induced by ME in histories (O/E).
- We are able to improve accuracy (MSEP).
- **We are able to improve discrimination (ROC-AUC) only to some degree.**

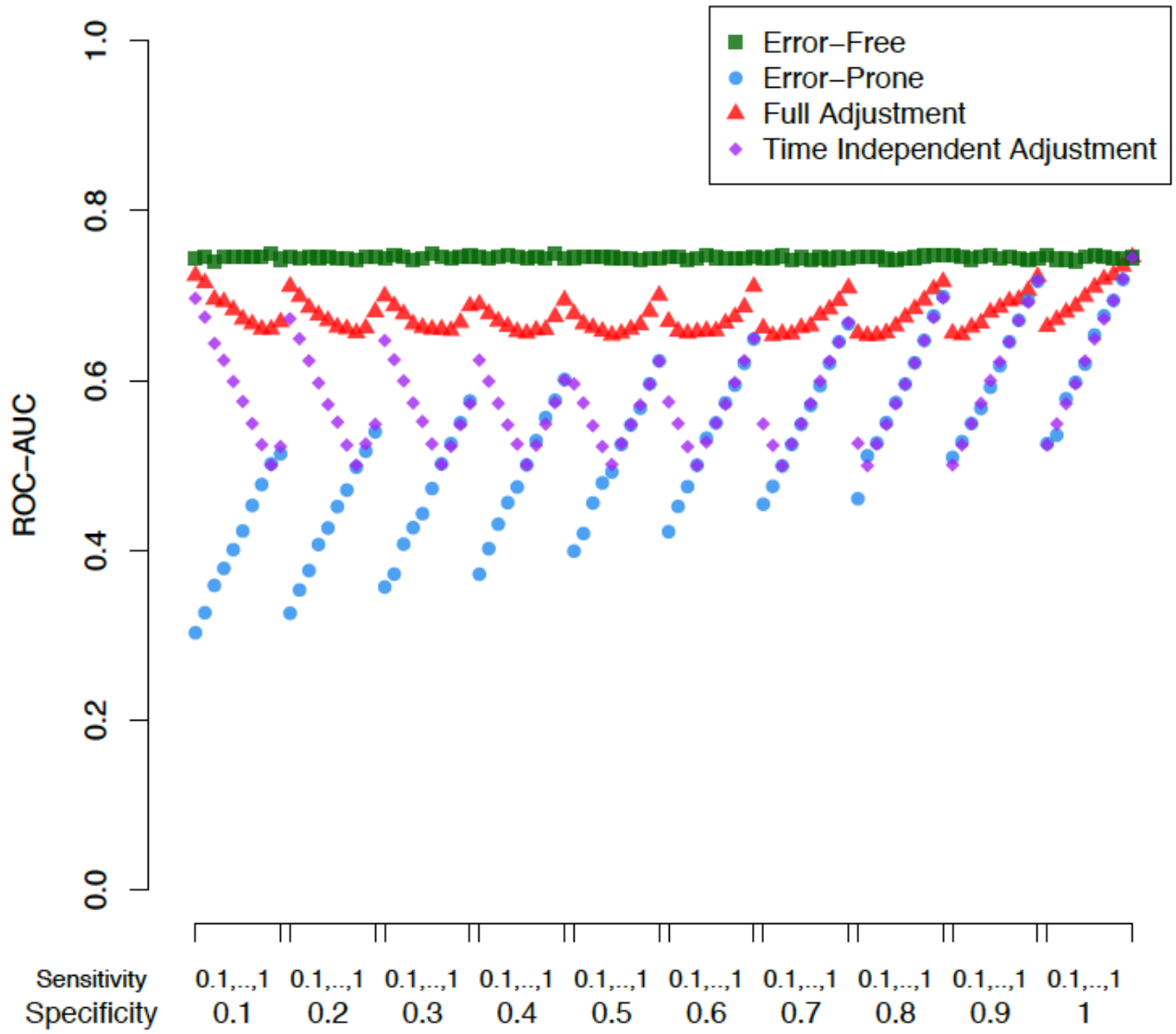
Survival prediction - summary of simulation results (multip.)



Survival prediction - summary of simulation results (multip.)



Survival prediction - summary of simulation results (multip.)



Concluding remarks

- ✓A non-parametric adjustment is provided, for measurement error in time to event predictor.
- ✓Ignoring the measurement error, provides miscalibrated models.
- ✓The proposed adjustment improves calibration and total accuracy.
- ✓The proposed method can be easily incorporated in BayesMendel R package for direct clinical use.

Model discrimination only partially improved.

END

Example – misreporting breast cancer

Counselees:

- Data from the Cancer Genetics Network (CGN) Model Evaluation Study, with known carrier status.
- 2038 families, 34310 relatives.
- 9.2% of the relatives have breast cancer.
- Only error-prone self-reported family history is available.

Validation data:

- Data from U of California at Irvine (UCI).
- 719 cancer affected counselees (breast, ovarian or colon cancer).
- 1521 female relatives, 19.3% with breast cancer.
- Error-prone and error-free family history are available.

Example – misreporting breast cancer

Log of O/E and 95% confidence intervals for being a BRCA carrier for counselees in CGN dataset, stratified by risk decile:

- **Transportability?**
- **Small sample?**

Very small improvement in Brier score and ROC-AUC.