A Scalable Approach to Measuring the Impact of Missing Data in Cancer Studies

For 2015 NCI Division of Cancer Control and Population Sciences New Grantees Workshop

Grant Number: R01CA178061

PI: Hui Xie

Co-Investigators: Donald Hedeker
Robin Mermelstein

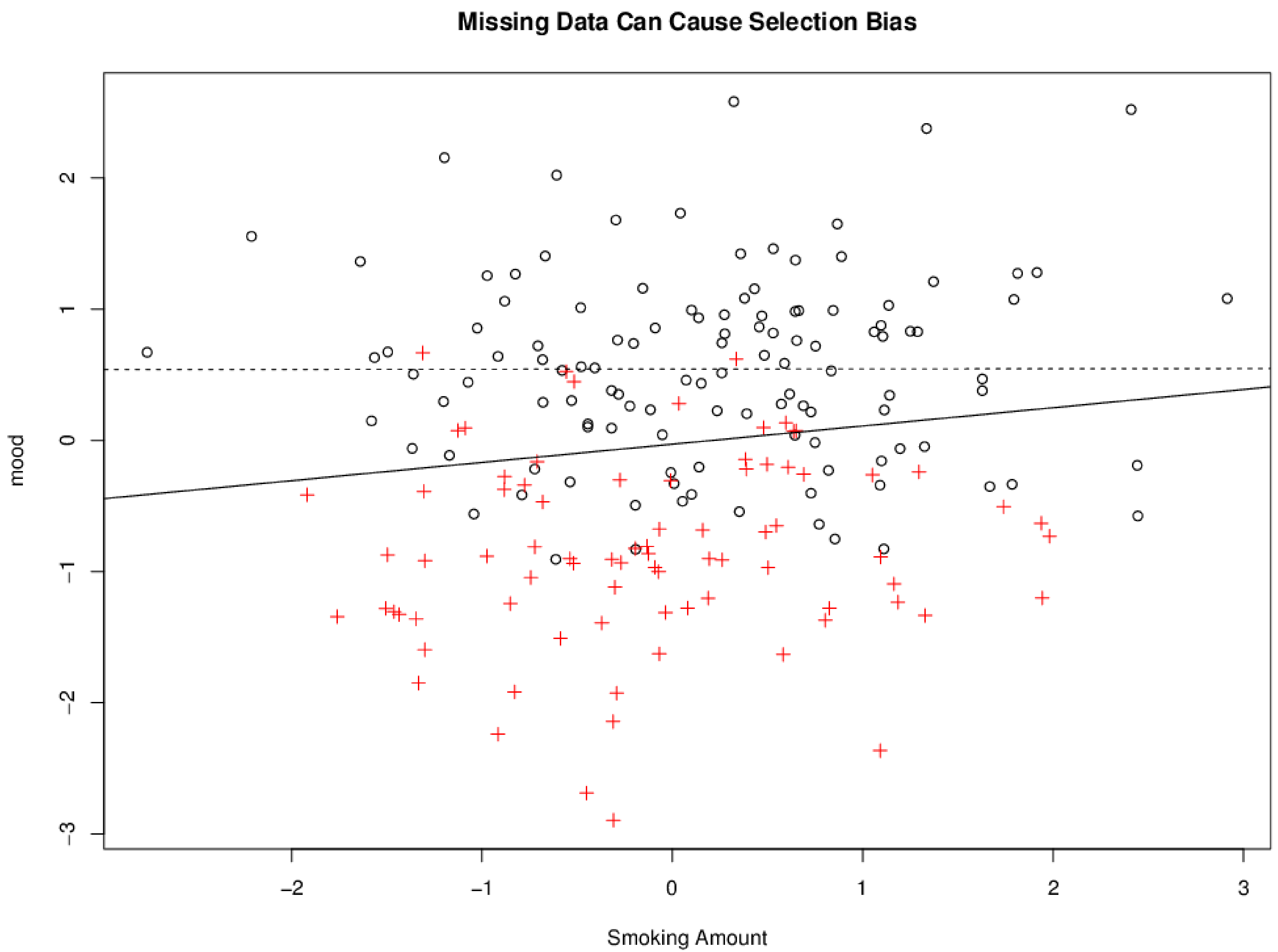
Research Assistant: Weihua Gao

Missing Data: Common in Cancer Studies

- Missing Data are ubiquitous in Cancer Studies, e.g.
  1. Dropout in longitudinal cancer epidemiological data
  2. Patient noncompliance in cancer clinical trials
  3. Caner Patients' online rating and satisfaction data
  4. Nonresponses in mobile health and other electronic capturing system.
- Missing Data issue does not go away with technology advancement.
- We consider the Ecological Momentary Assessment (EMA) data collection via hand-held palmtop computers from an NIH-funded longitudinal study of the natural history of smoking among adolescents (P01 CA098262, PI Mermelstein)

Benefits of EMA for Cancer Studies

- Utilize new technological devices in data collection.
- Random prompts fives times per day, asking about mood, activity, location, companionship and other behaviors.
- “Smoke” reports asked the same questions as random prompts plus smoking-related items.
- Real-time data collection captures subtle variations in mood.
- Increase data accuracy by minimizing recall bias.
- Provide richer data (many more observations per subject).
- Missing data are, however, unavoidable because of participants’ nonresponses to random prompts.
- The observed data may be a biased representation of the underlying distributions.


Missing Data Can Cause Selection Bias
mood
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
Smoking Amount

Understanding Missing Data Impact is Important

- In reporting results, reviewing proposals and making policy decisions, questions often arise such as "How reliable is my finding given attrition", "How will your results be affected when responders and nonresponders are different"....

- "Examining sensitivity to the assumptions about the missing data mechanism should be a mandatory component of reporting"

and

"The treatment of missing data in clinical trials, being a crucial issue, should have a higher priority for sponsors of statistical research, such as the National Institute of Health and National Science Foundation"

------- National Research Council (2008)

The challenges

- Option 1: Impute Missing observations with extreme values.

  Simple but ad-hoc; bound can be too wide to be useful

- Option 2: Check parameter changes by formally fitting a set of alternative nonignorable models.

For instance, for longitudinal data with dropout, one can check parameter changes under a set of Markov Transitional selection models with informative dropout, where dropout probability depends on the current outcome (potentially missing due to dropout) , the history of past outcomes and covariates, etc.

Principled but complex, can become intractable especially for rich datasets.

Limit the number, types and sizes of analyses that can be performed.

Hinder the use of principled statistical methods.

- Need a tractable and principled tool to quantify the effect of nonignorable missingness on standard analysis!

Our Solution: Simple yet Principled

- Approximate the parameter changes using first-order Taylor-series expansion.
- First-derivative, named as the Index of Sensitivity to nonignorability (ISNI, Troxel et al. 2004), measures the changing rate of estimates (i.e., Sensitivity) around the standard MAR model.
- No Need to fit any nonignorable models: Capable of Reducing computational time from hours or days to just a few seconds.
- Scalable to large datasets.
- Accurate for relevant and moderate nonignorability when a rich set of missingness predictors exisit (e.g., in longitudinal data or in a rich cross-sectional data)

A Smoking Cessation Example

- A randomized trial of Smoking-cessation study (Gruder et al. 1993).

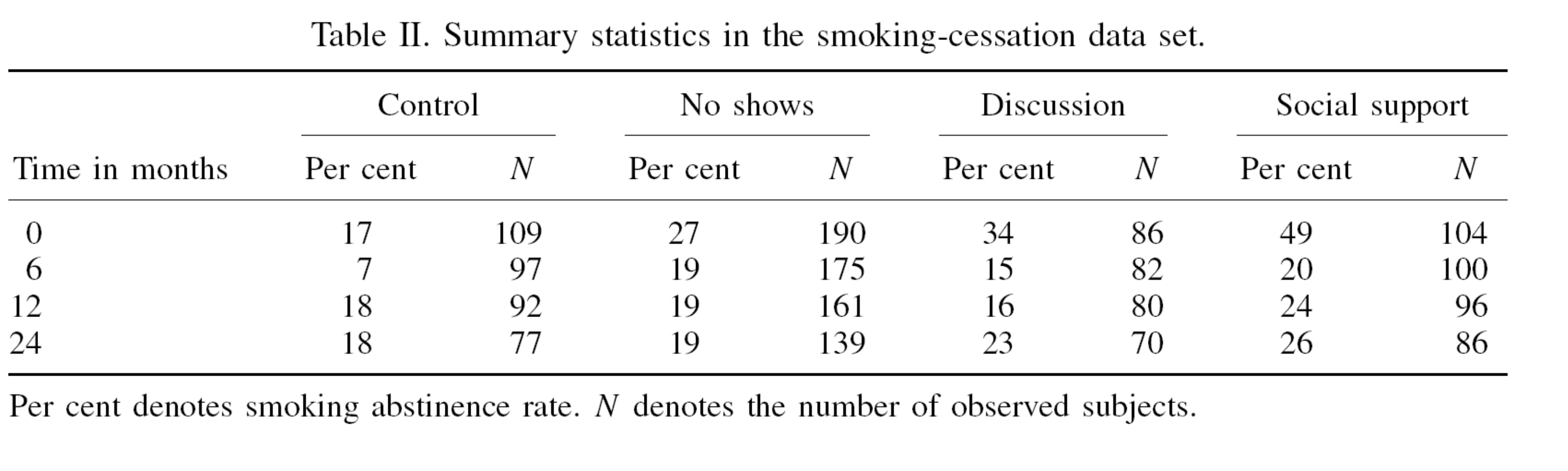


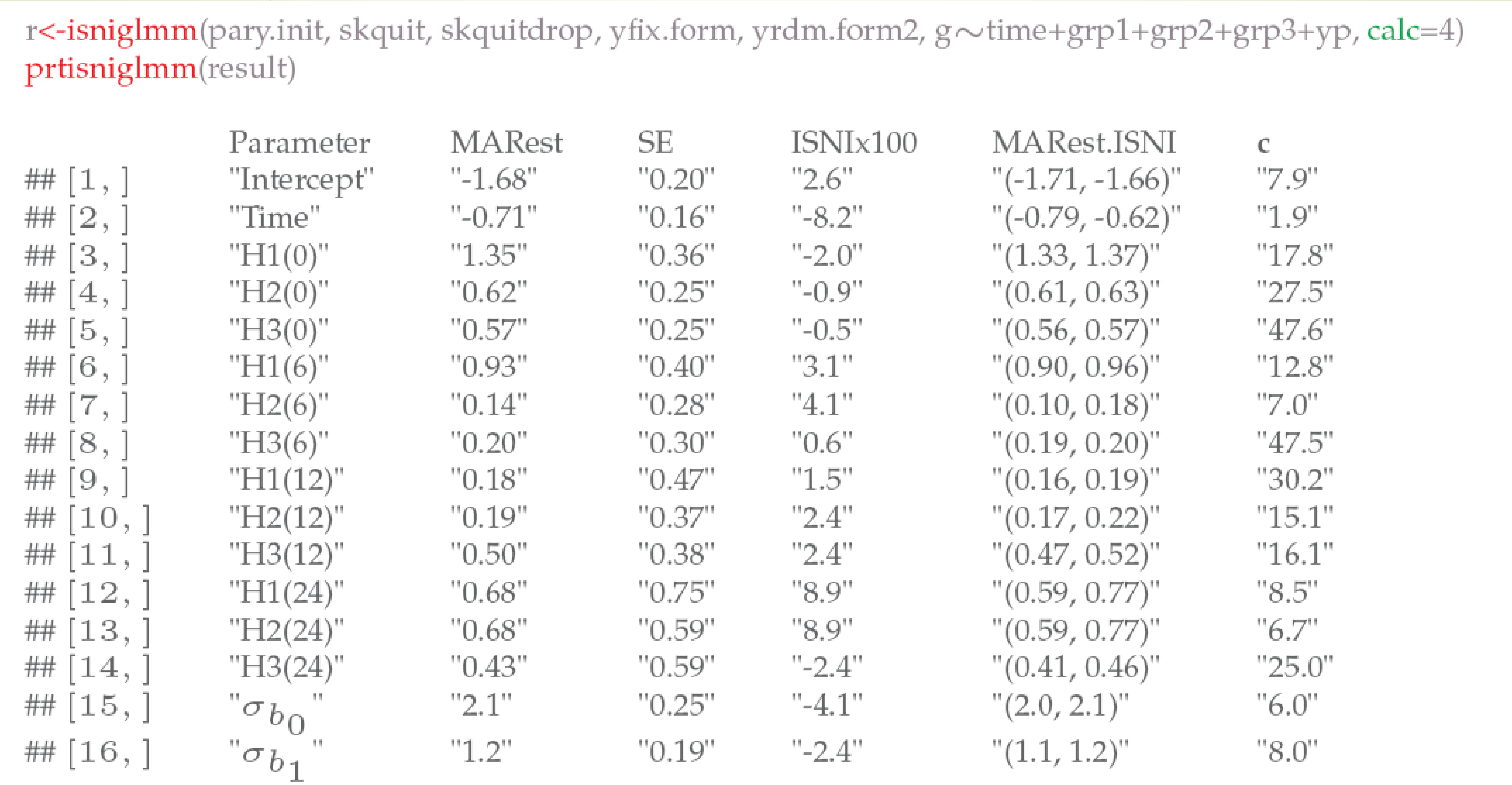
Table II. Summary statistics in the smoking-cessation data set.

	Control		No shows		Discussion		Social support	
Time in months	Per cent	*N*	Per cent	*N*	Per cent	*N*	Per cent	*N*
0	17	109	27	190	34	86	49	104
6	7	97	19	175	15	82	20	100
12	18	92	19	161	16	80	24	96
24	18	77	19	139	23	70	26	86

Per cent denotes smoking abstinence rate. *N* denotes the number of observed subjects.

- How trustworthy are the standard treatment effects comparisons assuming that dropout behavior can be fully explained by observed data elements?

Analysis using the New R Package ISNI



```
r<-isniglmm(pary.init, skquit, skquitdrop, yfix.form, yrdm.form2, g~time+grp1+grp2+grp3+yp, calc=4)
prtisniglmm(result)
```

```
              Parameter        MARest     SE        ISNIx100     MARest.ISNI          c
## [1, ]      "Intercept"      "-1.68"    "0.20"    "2.6"        "(-1.71, -1.66)"     "7.9"
## [2, ]      "Time"           "-0.71"    "0.16"    "-8.2"       "(-0.79, -0.62)"     "1.9"
## [3, ]      "H1(0)"          "1.35"     "0.36"    "-2.0"       "(1.33, 1.37)"       "17.8"
## [4, ]      "H2(0)"          "0.62"     "0.25"    "-0.9"       "(0.61, 0.63)"       "27.5"
## [5, ]      "H3(0)"          "0.57"     "0.25"    "-0.5"       "(0.56, 0.57)"       "47.6"
## [6, ]      "H1(6)"          "0.93"     "0.40"    "3.1"        "(0.90, 0.96)"       "12.8"
## [7, ]      "H2(6)"          "0.14"     "0.28"    "4.1"        "(0.10, 0.18)"       "7.0"
## [8, ]      "H3(6)"          "0.20"     "0.30"    "0.6"        "(0.19, 0.20)"       "47.5"
## [9, ]      "H1(12)"         "0.18"     "0.47"    "1.5"        "(0.16, 0.19)"       "30.2"
## [10, ]     "H2(12)"         "0.19"     "0.37"    "2.4"        "(0.17, 0.22)"       "15.1"
## [11, ]     "H3(12)"         "0.50"     "0.38"    "2.4"        "(0.47, 0.52)"       "16.1"
## [12, ]     "H1(24)"         "0.68"     "0.75"    "8.9"        "(0.59, 0.77)"       "8.5"
## [13, ]     "H2(24)"         "0.68"     "0.59"    "8.9"        "(0.59, 0.77)"       "6.7"
## [14, ]     "H3(24)"         "0.43"     "0.59"    "-2.4"       "(0.41, 0.46)"       "25.0"
## [15, ]     "$\sigma_{b_0}$"  "2.1"      "0.25"    "-4.1"       "(2.0, 2.1)"         "6.0"
## [16, ]     "$\sigma_{b_1}$"  "1.2"      "0.19"    "-2.4"       "(1.1, 1.2)"         "8.0"
```

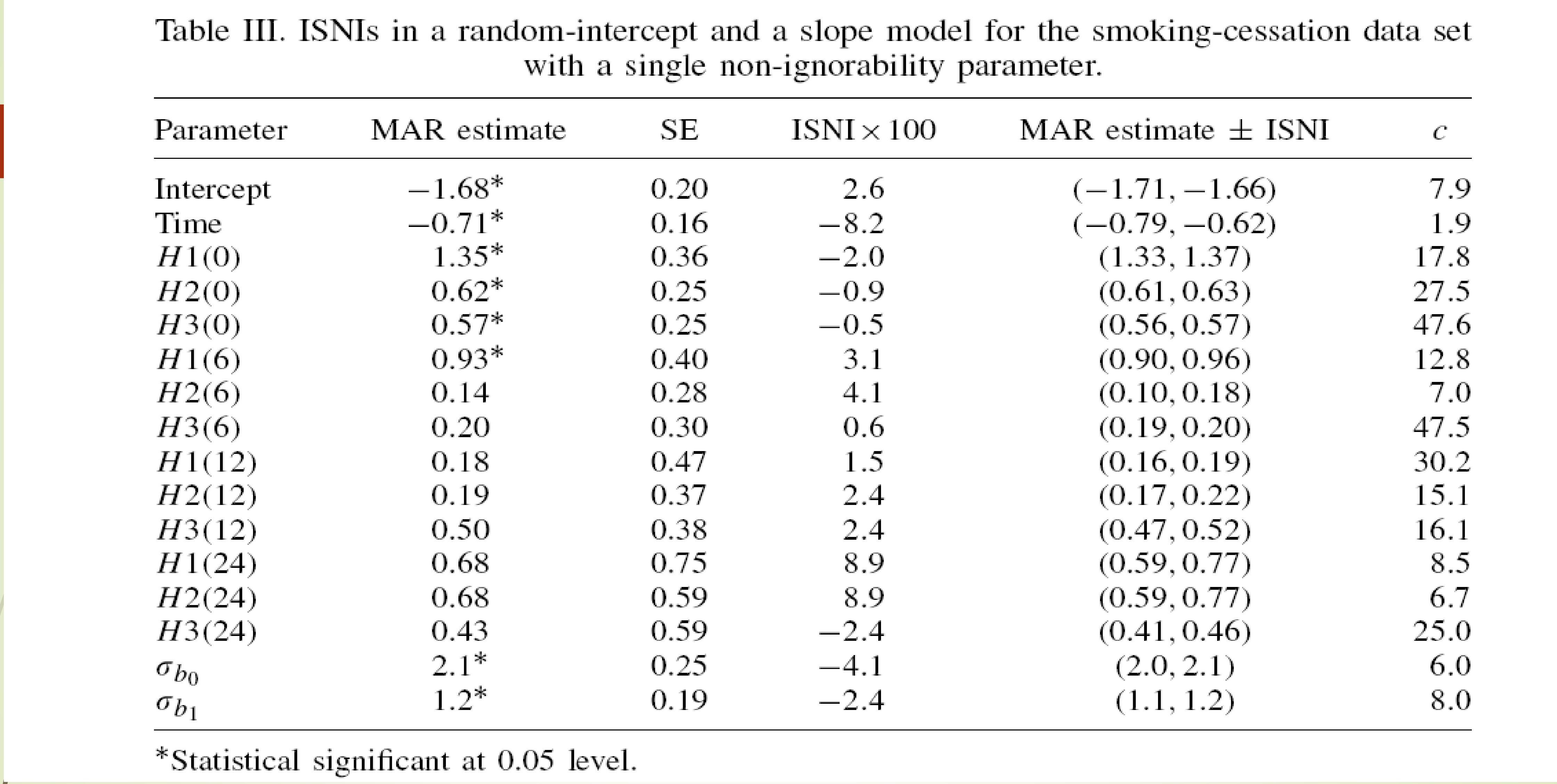


Table III. ISNIs in a random-intercept and a slope model for the smoking-cessation data set with a single non-ignorability parameter.

Parameter	MAR estimate	SE	ISNI × 100	MAR estimate ± ISNI	c
Intercept	−1.68*	0.20	2.6	(−1.71, −1.66)	7.9
Time	−0.71*	0.16	−8.2	(−0.79, −0.62)	1.9
$H1(0)$	1.35*	0.36	−2.0	(1.33, 1.37)	17.8
$H2(0)$	0.62*	0.25	−0.9	(0.61, 0.63)	27.5
$H3(0)$	0.57*	0.25	−0.5	(0.56, 0.57)	47.6
$H1(6)$	0.93*	0.40	3.1	(0.90, 0.96)	12.8
$H2(6)$	0.14	0.28	4.1	(0.10, 0.18)	7.0
$H3(6)$	0.20	0.30	0.6	(0.19, 0.20)	47.5
$H1(12)$	0.18	0.47	1.5	(0.16, 0.19)	30.2
$H2(12)$	0.19	0.37	2.4	(0.17, 0.22)	15.1
$H3(12)$	0.50	0.38	2.4	(0.47, 0.52)	16.1
$H1(24)$	0.68	0.75	8.9	(0.59, 0.77)	8.5
$H2(24)$	0.68	0.59	8.9	(0.59, 0.77)	6.7
$H3(24)$	0.43	0.59	−2.4	(0.41, 0.46)	25.0
σ_{b_0}	2.1*	0.25	−4.1	(2.0, 2.1)	6.0
σ_{b_1}	1.2*	0.19	−2.4	(1.1, 1.2)	8.0

*Statistical significant at 0.05 level.

- C: minimum (standardized) magnitude of nonignorable missingness needed to have important sensitivity. C=1 suggested as a cutoff value
- Sensitivity to nonignorable dropout increases from baseline to month 24.
- Treatment Effect comparisons are not sensitive to nonignorable dropout.

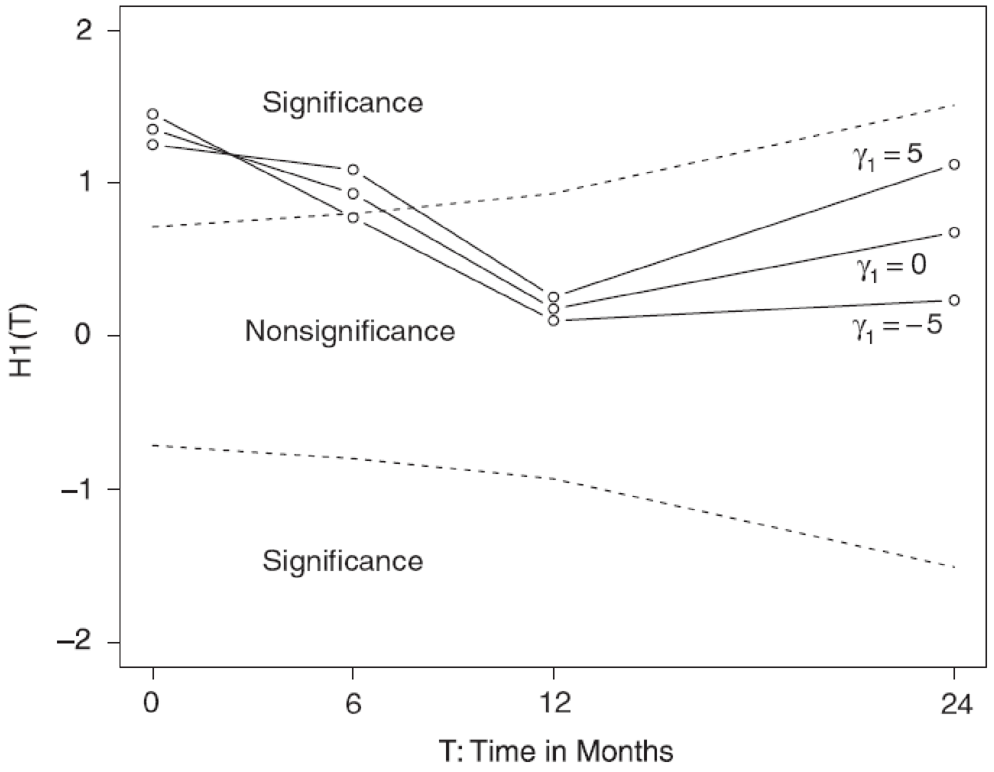


Figure 2. Sensitivity of the MAR estimates for the Helmert contrast $H1(T)$ at four visits. The solid line with $\gamma_1 = 0$ denotes the MAR estimates. The other two solid lines reflect the approximate change of the MLEs, approximated by the ISNI method, when γ_1 is varied to 5 and -5, respectively. Within the two dashed lines is the area of non-significance, where a test for $H1(T) = 0$ is not statistically significant.

Our R01 Study

- Overarching Goal: Derive new sensitivity indices for use with big data enabled by the advancement in information technology (e.g., electronic captured data, EMA data, mobile health data).
- Aim 1. Develop more general, flexible, robust and tractable methods and accessible software tailored for assessing the impact of missing data in EMA data.
- Aim 2. Examine the role of smoking in mood regulation in EMA studies of adolescents while accounting for the impact of nonignorable missingness using the methods and software developed in Aim 1.
- Current Focus

1. Method Development and Applications in EMA Data

A Scalable Approach to Measuring the Impact of Nonignorable Nonresponse with an EMA application. Technical Report 2015

2. Software Development in R and SAS

ISNI: A New R package to Measure the Impact of Nonignorability. R vignettes 2015.

Upon the Completion of the Project

- Provide insights about the role of smoking in mood regulation, adjusting for the impact of potential nonignorable missingness on these analyses using the EMA data.
- Provide researchers in EMA and other cancer-related fields with general and flexible methods and easy-to-use software programs to assess and improve the reliability of their study findings in the presence of missing data.
- Reporting results: Incorporate the sensitivity analysis results into the primary reporting.
- Screening sensitive data: Collect data on missing mechanism and do a more complex nonignorable modeling only if sensitivity is noticeable.

Question and Comments