smdi

An R package to perform structural missing data investigations for real-world evidence studies

Janick Weberpals, RPh, PhD
jweberpals@bwh.harvard.edu
Division of Pharmacoepidemiology and Pharmacoeconomics
Brigham and Women's Hospital
Harvard Medical School

October 26, 2023

Disclosures

Disclosures

- Janick Weberpals reports prior employment by Hoffmann-La Roche and previously held shares in Hoffmann-La Roche
- This project was supported by Task Order 75F40119F19002 under Master Agreement 75F40119D10037 from the U.S. Food and Drug Administration (FDA)

Background

Administrative insurance claims databases are increasingly linked to **electronic health records (EHR)** to improve confounding adjustment for variables which cannot be measured in administrative claims

Examples:

- Labs (HbA1c, LDL, etc.)
- Vitals (Blood pressure, BMI, etc.)
- Disease-specific data (cancer stage, biomarkers, etc.)
- Physician assessments (ECOG, etc.)
- Lifestyle factors (smoking, alcohol, etc.)

These covariates are often just partially observed for various reasons:

- Physician did not perform/order a certain test
- Certain measurements are just collected for particularly sick patients
- Information is 'hiding' in unstructured records, e.g. clinical notes

Knowledge gaps and objectives

Missing data in EHR confounding factors are frequent

Two common missing data taxonomies

- **Mechanisms**: Missing completely at random (MCAR), at random (MAR) and not at random (MNAR)
- **Patterns**: Monotone, Non-monotone

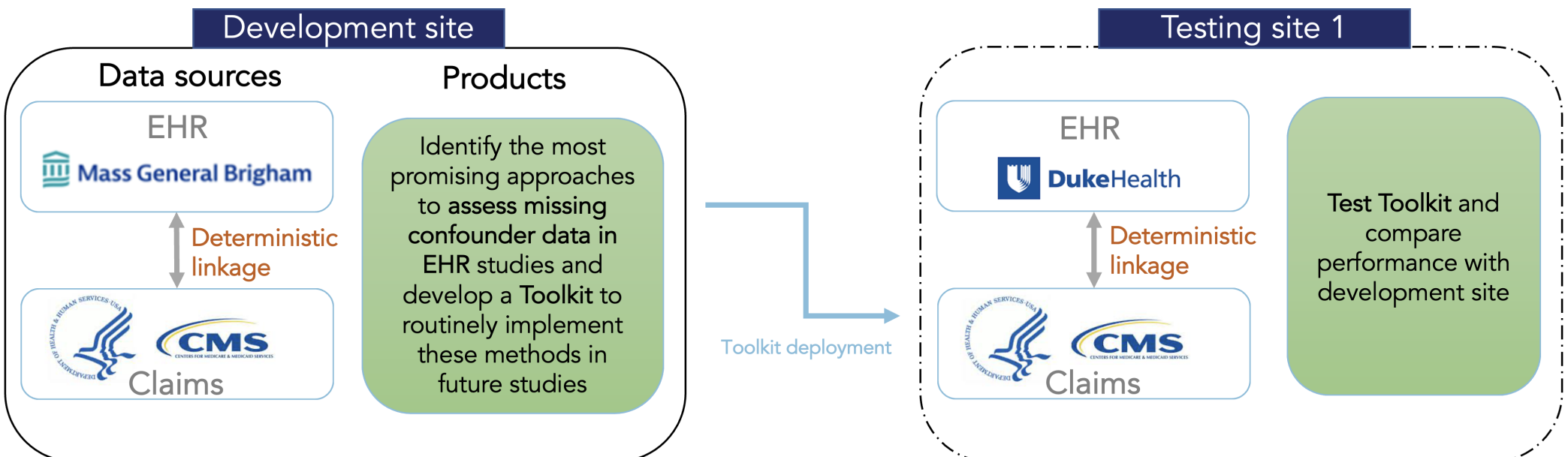
Unresolved challenges for `causal inference`:

- In an empirical study, it is usually unclear which of the missing data patterns and mechanisms are dominating.
- What covariate relationships exist and are partially observed covariates recoverable **in high-dimensional covariate spaces (e.g., database linkages)**?

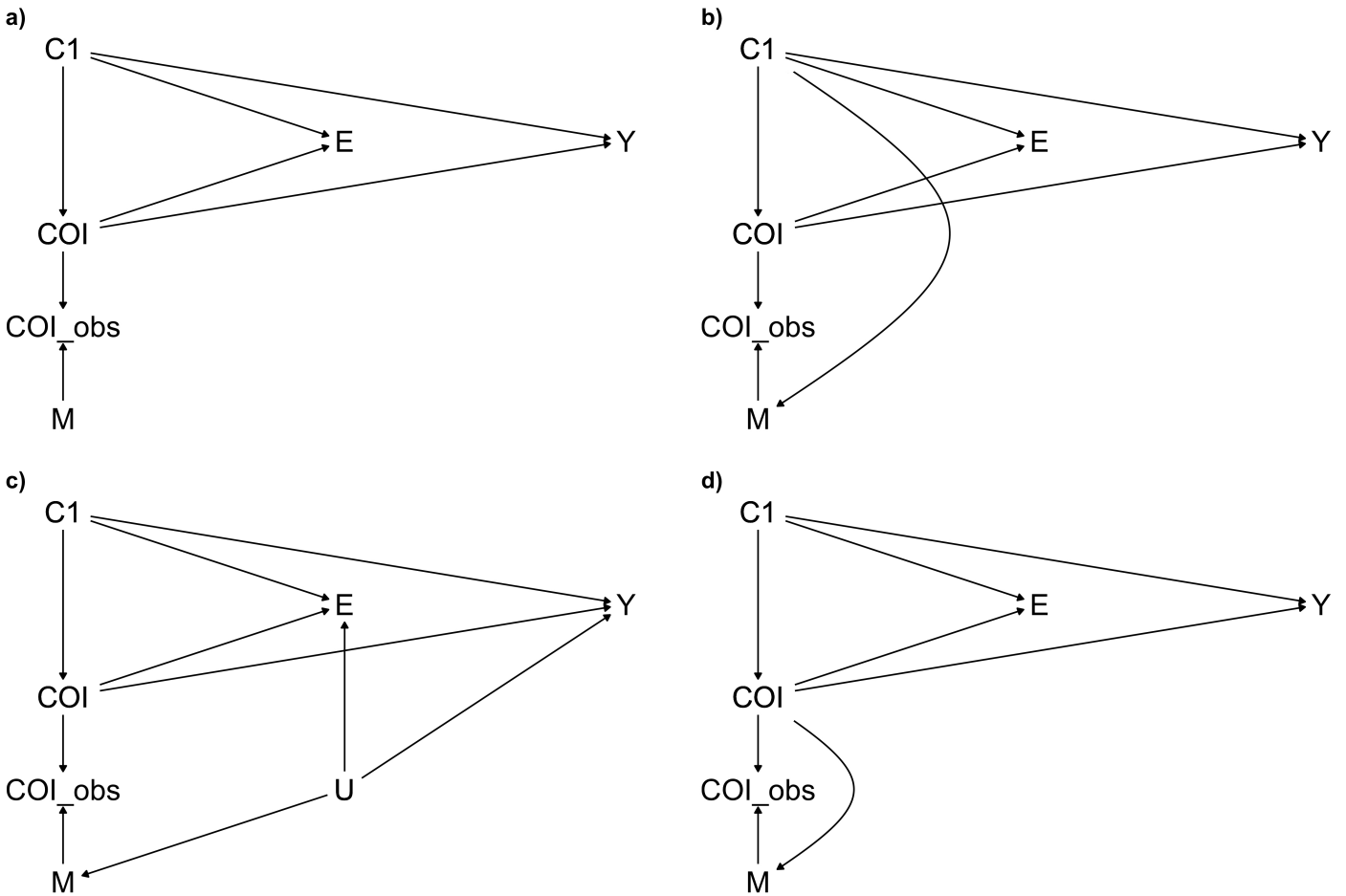
Objectives

Objectives of the Sentinel Innovation Center Causal Inference Workstream

- Develop a framework and tools to assess the structure of missing data processes in EHR studies
- Connect this with the most appropriate analytical approach, followed by sensitivity analyses
- **Develop an R package to implement framework and missing data investigations on a routine basis**



Assumed missingness structures



Causal diagrams/M-graphs[1,2] provide a more natural way to understand the assumptions regarding missing (confounder) data for a given research question, Legend: a) Missing completely at random (MCAR), b) Missing at random (MAR), c) Missing not at random 1 (MNAR unmeasured), d) Missing not at random 2 (MNAR value), Notation: E = Exposure, Y = Outcome, C1 = Fully observed confounders, C = Confounder of interest, C_obs = Observed portion of C, M = Missingness indicator

Missing data diagnostics

	Group 1 Diagnostics		Group 2 Diagnostics	Group 3 Diagnostics
	Median Absolute standardized mean difference (ASMD)	P-value Hoteling/Little	AUC (area under the receiver operating characteristic curve)	Log HR (missingness indicator)
Purpose	Comparison of distributions of observed covariates between patients with vs w/o observed value of the partially observed confounder		Assessing the ability to predict confounder missingness based on observed covariates	Check whether confounder missingness is associated with the outcome (differential missingness)
Example value	ASMD = 0.1	p-value <0.001	AUC = 0.5	log HR = 0.1 (0.05 to 0.2)
Interpretation	<0.1*: no imbalances in observed patient characteristics; missingness may be likely completely at random or not at random (~MCAR, ~MNAR). >0.1*: imbalances in observed patient characteristics; missingness may be likely at random (~MAR). * Equivalent to propensity score-based balance measures (Austin PC, Multivariate Behavioral Research, 46:3, 399-424 [2011])	High test statistics and low p-values indicate differences in baseline covariate distributions and null hypothesis would be rejected (~MAR). Hotelling H. Ann Math Stat. 2(3):360-378. (1931) & Little RJA. J Am Stat Assoc. 83(404):1198-1202. doi:10.2307/2290157 (1988)	AUC values ~ 0.5 indicate completely random or not at random prediction (~MCAR, ~MNAR). Values meaningfully above 0.5 indicate stronger relationships between covariates and missingness (~MAR).	No association in either univariate or adjusted model and no meaningful difference in the log HR after full adjustment (~MCAR). Association in univariate but not fully adjusted model (~MAR). Meaningful difference in the log HR also after full adjustment (~MNAR).

Plasmode simulation - results

Observations

- Large scale simulation revealed characteristic patterns of the diagnostic parameters matched to missing data structure
- The observed diagnostic pattern of a specific study will give insights into the likelihood of underlying EHR missingness structures

	Group 1 Diagnostics		Group 2 Diagnostics	Group 3 Diagnostics	
Expected parameter constellations	**ASMD** (Absolute standardized mean difference)	**P-value** Hoteling/Little	**AUC** (are under the receiver operating curve)	**Log HR** (crude)	**Log HR** (adjusted)
MCAR	0.05	0.5	0.50	-0.01	0.00
MAR	0.20	<.001	0.58	0.53	0.00
$MNAR_{unmeasured}$	0.09	0.02	0.54	0.43	0.31
$MNAR_{value}$	0.06	0.10	0.53	0.04	0.10

Plasmode simulation results averaged across all scenarios and simulated datasets.

Plasmode simulation - results

	Group 1 Diagnostics		Group 2 Diagnostics	Group 3 Diagnostics	
Expected parameter constellations	**ASMD** (Absolute standardized mean difference)	**P-value** Hoteling/Little	**AUC** (are under the receiver operating curve)	**Log HR** (crude)	**Log HR** (adjusted)
MCAR	0.05	0.5	0.50	-0.01	0.00
MAR	0.20	<.001	0.58	0.53	0.00
$MNAR_{unmeasured}$	0.09	0.02	0.54	0.43	0.31
$MNAR_{value}$	0.06	0.10	0.53	0.04	0.10

Let's have a look at some EHR examples:

Covariate	**ASMD** (min to max)	**P-value**	**AUC**	**Log HR** (crude, 95% CI)	**Log HR** (adjusted, 95% CI)
EGFR (cancer biomarker)	0.24 (0.01 to 0.49)	<.001	0.63	0.06 (-0.03 to 0.15)	-0.01 (-0.10 to 0.09)

The observed diagnostic pattern of a specific study will give insights into the likelihood of underlying missingness structures

Plasmode simulation - results

	Group 1 Diagnostics		Group 2 Diagnostics	Group 3 Diagnostics	
Expected parameter constellations	**ASMD** (Absolute standardized mean difference)	**P-value** Hoteling/Little	**AUC** (are under the receiver operating curve)	**Log HR** (crude)	**Log HR** (adjusted)
MCAR	0.05	0.5	0.50	-0.01	0.00
MAR	0.20	<.001	0.58	0.53	0.00
$MNAR_{unmeasured}$	0.09	0.02	0.54	0.43	0.31
$MNAR_{value}$	0.06	0.10	0.53	0.04	0.10

Let's have a look at some EHR examples:

Covariate	**ASMD** (min to max)	**P-value**	**AUC**	**Log HR** (crude, 95% CI)	**Log HR** (adjusted, 95% CI)
EGFR (cancer biomarker)	0.24 (0.01 to 0.49)	<.001	0.63	0.06 (-0.03 to 0.15)	-0.01 (-0.10 to 0.09)
ECOG (performance status)	0.03 (0.00 to 0.07)	0.78	0.51	-0.06 (-0.16 to 0.03)	-0.06 (-0.16 to 0.03)

The observed diagnostic pattern of a specific study will give insights into the likelihood of underlying missingness structures

The whole game - `smdi` workflow to perform routine missing data diagnostics

exposure	age_num	female_cat	smoking_cat	physical_cat	alk_cat	histology_cat	ses_cat	copd_cat	eventtime	status	ecog_cat	egfr_cat	pdl1_num
1	35.24	1	1	0	0	1	2_middle	1	5.000000000	0	1	*NA*	45.03
1	51.18	0	1	1	0	1	3_high	0	4.754220474	1	*NA*	0	*NA*
0	88.17	0	0	0	0	0	2_middle	1	0.253391563	1	0	1	41.74
1	50.79	0	1	0	0	0	2_middle	1	5.000000000	0	1	*NA*	45.51
1	40.52	0	1	0	0	0	2_middle	1	5.000000000	0	*NA*	1	31.28

Dataframe with one row per patient and relevant variables as columns
(exposure, outcome, covariates, partially observed covariates)

Descriptives And Pattern Diagnostics

Which covariates exhibit missingness?

- *smdi_check_covar()*
- *smdi_na_indicator()*

Summarize and visualize missingness:

- *smdi_summarize()*
- *smdi_vis()*

Identify patterns visually:*

- *gg_miss_upset()*
- *md_pattern()*

Inferential Three Group Diagnostics

Group 1 Diagnostics

- *smdi_amsd()*
- *smdi_hotelling()*
- *smdi_little()*

Group 2 Diagnostics

- *smdi_rf()*

Group 3 Diagnostics

- *smdi_outcome()*

Group 1-3 Diagnostics

- *smdi_diagnose()*
- *smdi_style_gt()*

*If pattern seems **non-monotone** → run diagnostics on all partially observed covariates jointly, if **monotone** consider running diagnostics on each partially observed covariate individually*

Suggested `smdi` workflow.

smdi bundled datasets

- The smdi package comes with two exemplary **simulated** datasets:
  - smdi_data (includes some partially observed covariates)
  - smdi_data_complete (complete dataset if you prefer to introduce NA yourself)

```
1 smdi_data |>
2   glimpse()
```

```
Rows: 2,500
Columns: 14
$ exposure      <int> 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0,…
$ age_num       <dbl> 35.24, 51.18, 88.17, 50.79, 40.52, 64.57, 73.58, 42.38, …
$ female_cat    <fct> 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1,…
$ smoking_cat   <fct> 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1,…
$ physical_cat  <fct> 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0,…
$ alk_cat       <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ histology_cat <fct> 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0,…
$ ses_cat       <fct> 2_middle, 3_high, 2_middle, 2_middle, 2_middle, 2_middle…
$ copd_cat      <fct> 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1,…
$ eventtime     <dbl> 5.000000000, 4.754220474, 0.253391563, 5.000000000, 5.00…
$ status        <int> 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1,…
$ ecog_cat      <fct> 1, NA, 0, 1, NA, 0, 1, 0, 1, NA, 1, NA, NA, 1, 1, 0, 1, …
$ egfr_cat      <fct> NA, 0, 1, NA, 1, NA, NA, 0, NA, 0, 1, NA, 0, NA, NA, 0, …
$ pdl1_num      <dbl> 45.03, NA, 41.74, 45.51, 31.28, NA, 47.28, 37.28, 46.47,…
```

Descriptives

- Let's start with some light descriptives
- All `smdi` functions automatically include all variables with at least one missing value (default)
- Investigator-specified variables can be selected via the `covar` parameter

```
smdi_data |>
  smdi_summarize()
```

```
# A tibble: 3 × 4
  covariate n_miss prop_miss prop_miss_label
  <chr>      <int>     <dbl> <chr>
1 egfr_cat    1015      40.6 40.60%
2 ecog_cat     899      36.0 35.96%
3 pdl1_num     517      20.7 20.68%
```

Descriptives - pattern

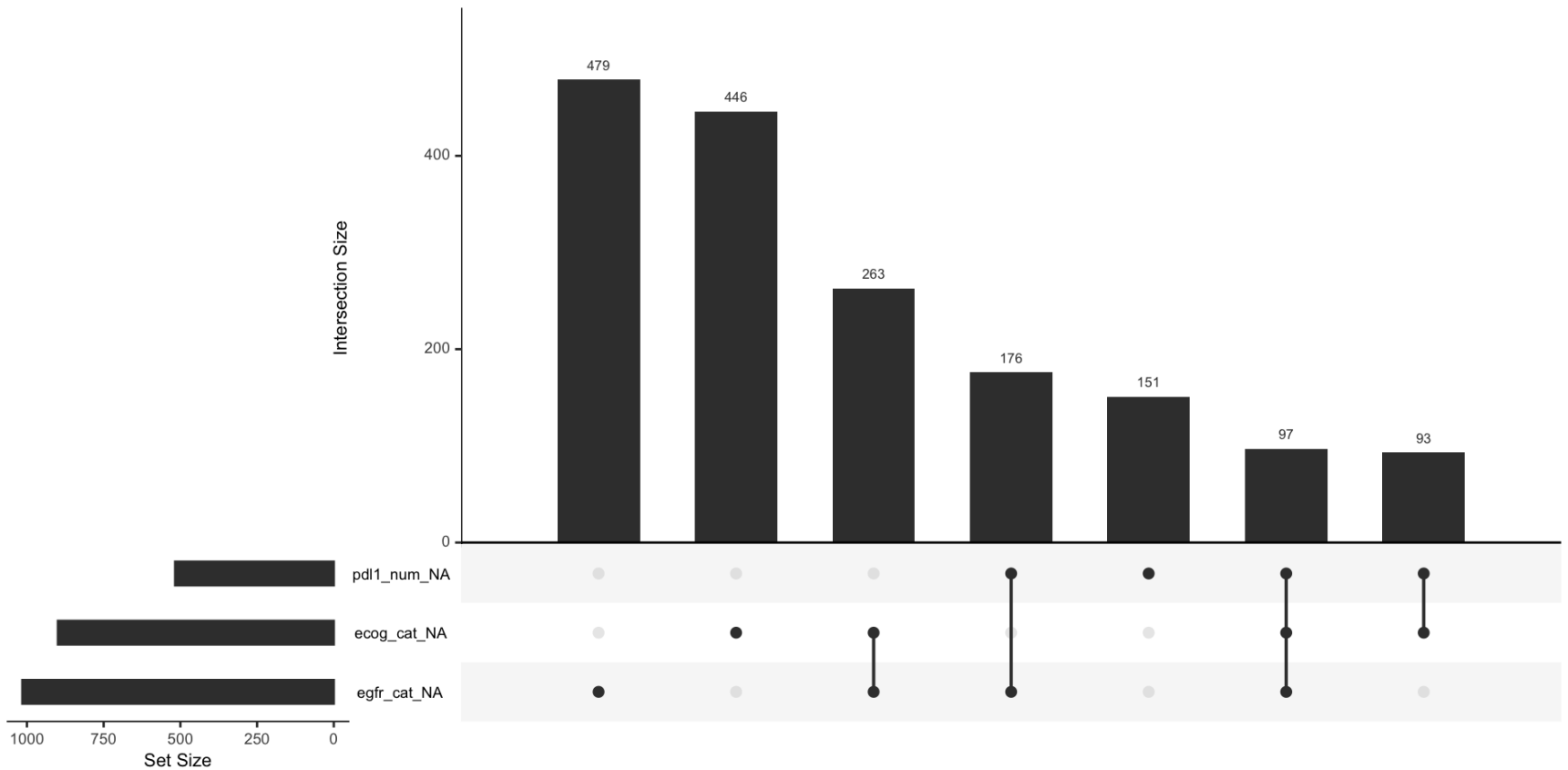
`smdi` uses a *re-export* of the **naniar**[3] `gg_miss_upset` and **mice**[4] `md.pattern` functions to investigate potentially underlying **missing data patterns**

> **Note**
>
> *Monotone and non-monotone (or general)*. A missing data pattern is said to be *monotone* if the variables \(Y_j\) can be ordered such that if \(Y_j\) is missing then all variables \(Y_k\) with \(k>j\) are also missing. This occurs, for example, in longitudinal studies with drop-out. If the pattern is not monotone, it is called *non-monotone* or *general*.[4]

Descriptives - pattern

`smdi` uses a *re-export* of the **naniar**[3] `gg_miss_upset` function to investigate potentially underlying **missing data patterns**

```
smdi_data |>
  gg_miss_upset()
```



smdi_asmd

> Group 1 diagnostics: Differences in covariate distributions

```
1  asmd <- smdi_asmd(data = smdi_data, median = TRUE, includeNA = FALSE)
2  asmd
```

```
# A tibble: 3 × 4
  covariate asmd_median asmd_min asmd_max
* <chr>     <chr>       <chr>    <chr>
1 ecog_cat  0.029       0.003    0.071
2 egfr_cat  0.243       0.010    0.485
3 pdl1_num  0.062       0.019    0.338
```

smdi_asmd

> Group 1 diagnostics: Differences in covariate distributions

```
asmd <- smdi_asmd(data = smdi_data, median = TRUE, includeNA = FALSE)
asmd
```

```
# A tibble: 3 × 4
  covariate asmd_median asmd_min asmd_max
* <chr>     <chr>       <chr>    <chr>
1 ecog_cat  0.029       0.003    0.071
2 egfr_cat  0.243       0.010    0.485
3 pdl1_num  0.062       0.019    0.338
```

The output returns an *asmd* object with much more information than what is captured in the S3 generic *print* output, e.g. a complete '*Table 1*' that displays the covariate distributions of patients:

```
head(asmd$pdl1_num$asmd_table1)
```

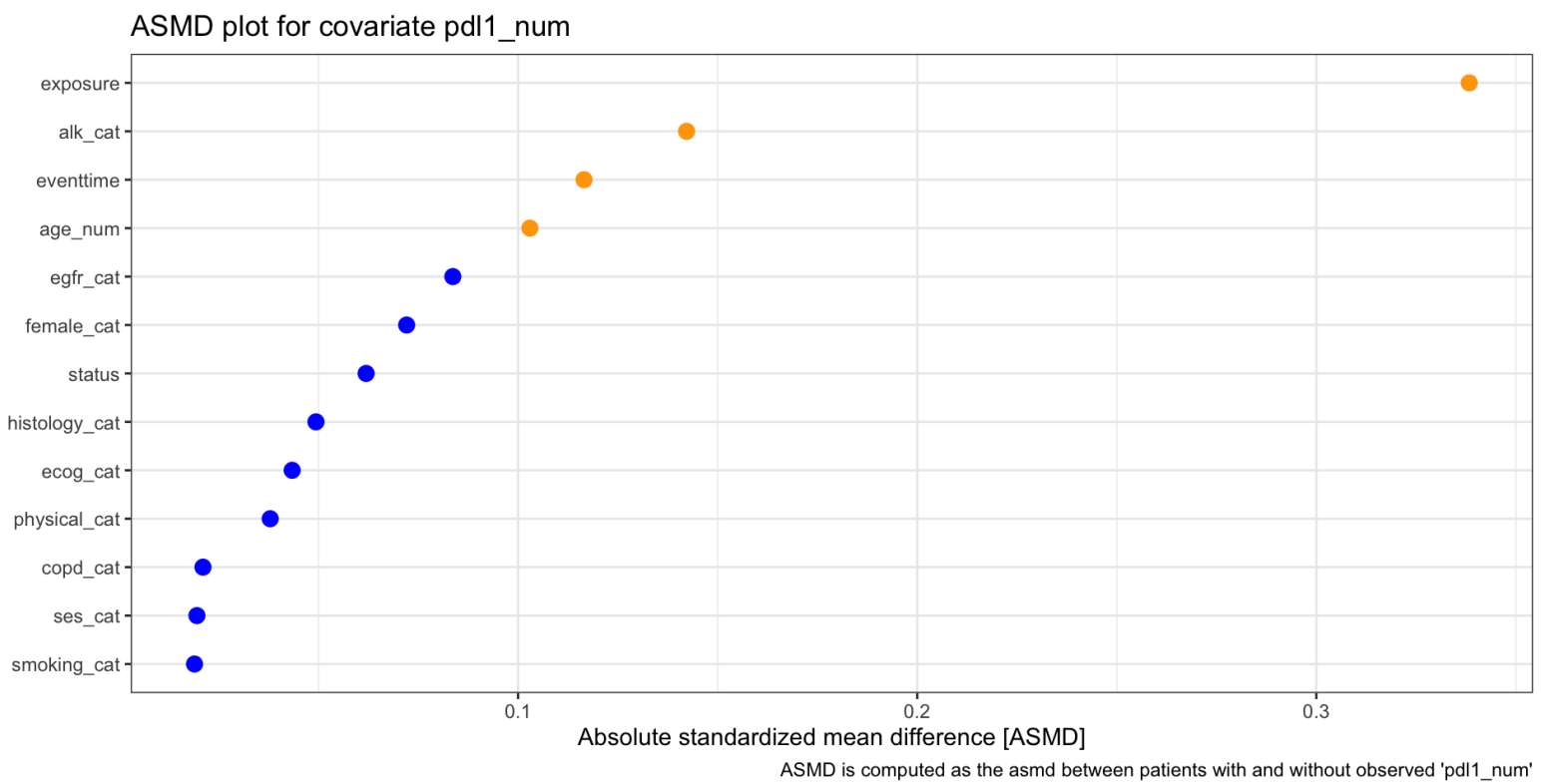
```
                     Stratified by pdl1_num_NA
                      0               1               p        test SMD
n                     " 1983"         "  517"         ""       ""   ""
exposure (mean (SD)) " 0.43 (0.50)"  " 0.27 (0.45)"  "<0.001" ""   " 0.338"
age_num (mean (SD))  "60.60 (14.04)" "62.07 (14.47)" " 0.036" ""   " 0.103"
female_cat = 1 (%)    "  717 (36.2) " "  205 (39.7) " " 0.157" ""   " 0.072"
smoking_cat = 1 (%)   "  990 (49.9) " "  263 (50.9) " " 0.739" ""   " 0.019"
physical_cat = 1 (%) "  707 (35.7) " "  175 (33.8) " " 0.476" ""   " 0.038"
```

smdi_asmd

> Group 1 diagnostics: Differences in covariate distributions

Investigators can also inspect standardized mean differences[5] by covariate in detail:

```
1 asmd$pdl1_num$asmd_plot
```



smdi_hotelling

> Group 1 diagnostics: Differences in covariate distributions

Hotelling's[6] multivariate t-test examines differences in covariate distributions conditional on having an observed covariate value or not. Rejection of \(H0\) would indicate significant differences between these patient strata.

```
1  smdi_hotelling(data = smdi_data)
```

```
  covariate hotteling_p
1  ecog_cat       0.783
2  egfr_cat       <.001
3  pdl1_num       <.001
```

smdi_little

> Group 1 diagnostics: Differences in covariate distributions

Little's[7] chi-square test takes into account possible patterns of missingness **across all variables** in the dataset. A high test statistics and low p-value (rejection of \(H0\)) would indicate that the **global** missing data generating mechanism is not completely at random.

```
1  smdi_little(data = smdi_data)
```

```
$statistic
[1] 801.0009

$df
[1] 86

$p.value
[1] 0

$missing.patterns
[1] 8

attr(,"class")
[1] "little"
attr(,"row.names")
[1] 1
```

smdi_rf

> Group 2 diagnostics: Ability to predict missingness

The `smdi_rf` function trains and fits a random forest model to assess the ability to predict missingness for the specified covariate(s).[8]

```
auc <- smdi_rf(data = smdi_data, train_test_ratio = c(.7, .3), set_seed = 42, n_cores = 3)
auc
```

```
# A tibble: 3 × 2
  covariate rf_auc
* <chr>     <chr>
1 ecog_cat  0.510
2 egfr_cat  0.629
3 pdl1_num  0.516
```

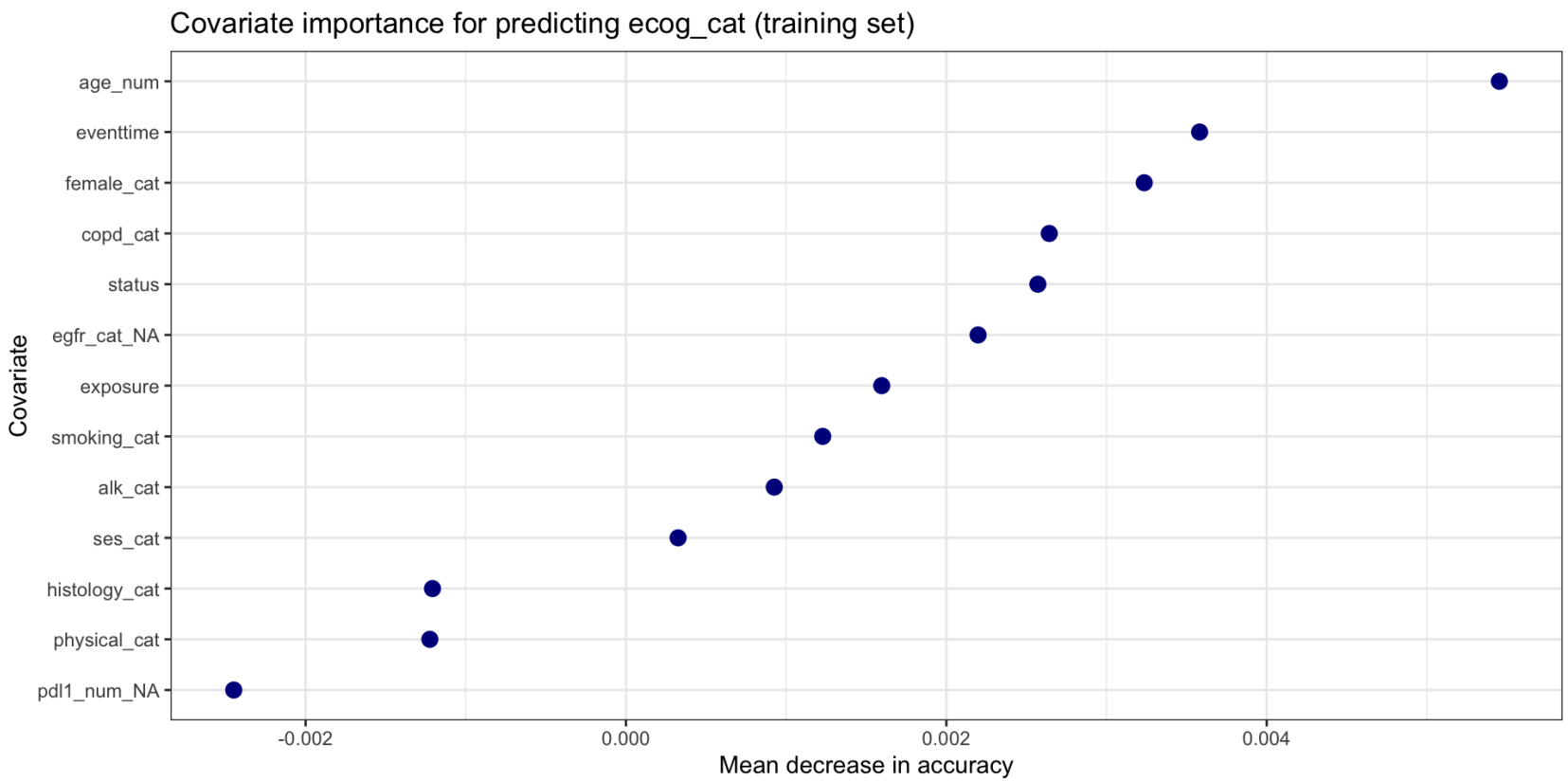
Parallelization

Depending on the amount of data (sample size x covariates), the computation of the function can take some minutes. To speed this up, investigators can parallelize the computation using `n_cores` (UNIX only).

smdi_rf

The resulting `smdi_rf` object provides the flexibility to investigate the covariate importance of predictors which can give important hints on the potentially underlying missing data generating mechanism.

```
1  auc$ecog_cat$rf_plot
```



smdi_outcome

Group 3 diagnostic focuses on assessing the association between the missing indicator of the partially observed covariate and the outcome under study (is the missingness differential?).

```
outcome <- smdi_outcome(
  data = smdi_data,
  model = "cox",
  form_lhs = "Surv(eventtime, status)",
  exponentiated = FALSE
  )

outcome
```

```
# A tibble: 3 × 3
  covariate estimate_univariate        estimate_adjusted
  <chr>     <glue>                     <glue>
1 ecog_cat  -0.06 (95% CI -0.16, 0.03) -0.06 (95% CI -0.16, 0.03)
2 egfr_cat  0.06 (95% CI -0.03, 0.15)  -0.01 (95% CI -0.10, 0.09)
3 pdl1_num  0.12 (95% CI 0.01, 0.23)   0.11 (95% CI -0.00, 0.22)
```

Supported regression types

Currently, the main types of outcome regressions are supported, namely *logistic (`glm`)*, *linear (`lm`)* and *Cox proportional hazards (`survival`)* models are supported and need to be specified using the model and form_lhs.

smdi_diagnose

One function to rule them all: `smdi_diagnose`

- Wrapper around all of the aforementioned functions
- Input parameters correspond to parameters of the individual functions

Let's take a look at a most minimal example

```
diagnostics <- smdi_diagnose(
  data = smdi_data,
  model = "cox",
  form_lhs = "Surv(eventtime, status)",
  n_cores = 3
  )

diagnostics
```

```
smdi summary table:
# A tibble: 3 × 6
  covariate asmd_median_min_max  hotteling_p rf_auc estimate_univariate
  <chr>     <chr>                <chr>       <chr>  <glue>
1 ecog_cat  0.029 (0.003, 0.071) 0.783       0.510  -0.06 (95% CI -0.16, 0.03)
2 egfr_cat  0.243 (0.010, 0.485) <.001       0.629  0.06 (95% CI -0.03, 0.15)
3 pdl1_num  0.062 (0.019, 0.338) <.001       0.516  0.12 (95% CI 0.01, 0.23)
# ℹ 1 more variable: estimate_adjusted <glue>

p_little: <.001
```

smdi_diagnose

Output is a list that resembles all three group diagnostics validated in the plasmode simulation study…

Covariate-specific table:

```
diagnostics$smdi_tbl
```

```
# A tibble: 3 × 6
  covariate asmd_median_min_max  hotteling_p rf_auc estimate_univariate
  <chr>     <chr>                <chr>       <chr>  <glue>
1 ecog_cat  0.029 (0.003, 0.071) 0.783       0.510  -0.06 (95% CI -0.16, 0.03)
2 egfr_cat  0.243 (0.010, 0.485) <.001       0.629  0.06 (95% CI -0.03, 0.15)
3 pdl1_num  0.062 (0.019, 0.338) <.001       0.516  0.12 (95% CI 0.01, 0.23)
# ℹ 1 more variable: estimate_adjusted <glue>
```

Global Little's test p-value:

```
diagnostics$p_little
```

```
p_little: <.001
```

smdi_style_gt

`smdi_style_gt` takes an object of class *smdi* (i.e., the output of `smdi_diagnose`) and formats it into a **publication-ready** `gt` table:

```
diagnostics |>
  smdi_style_gt(font_size = 18, tbl_width = 1000)
```

Covariate	ASMD (min/max)[1]	p Hotelling[1]	AUC[2]	beta univariate (95% CI)[3]	beta (95% CI)[3]
ecog_cat	0.029 (0.003, 0.071)	0.783	0.510	-0.06 (95% CI -0.16, 0.03)	-0.06 (95% CI -0.16, 0.03)
egfr_cat	0.243 (0.010, 0.485)	<.001	0.629	0.06 (95% CI -0.03, 0.15)	-0.01 (95% CI -0.10, 0.09)
pdl1_num	0.062 (0.019, 0.338)	<.001	0.516	0.12 (95% CI 0.01, 0.23)	0.11 (95% CI -0.00, 0.22)

p little: <.001, Abbreviations: ASMD = Median absolute standardized mean difference across all covariates, AUC = Area under the curve, beta = beta coefficient, CI = Confidence interval, max = Maximum, min = Minimum

[1] Group 1 diagnostic: Differences in patient characteristics between patients with and without covariate

[2] Group 2 diagnostic: Ability to predict missingness

[3] Group 3 diagnostic: Assessment if missingness is associated with the outcome (univariate, adjusted)

smdi_style_gt

Since `smdi_style_gt` transforms the *smdi* object into an object of class *gt_tbl*, an investigator can also take advantage of all of the `gt` package perks, e.g. exporting the table in different formats, e.g. .docx, .rtf, .pdf, etc.:

```
gtsave(
  data = smdi_style_gt(diagnostics),
  filename = "smdi_table.docx", # name of the final file and file type (e.g., .docx)
  path = "." # path where the file should be stored
  )
```

Test it out yourself

```r
# CRAN (current version: 0.2.2)
install.packages("smdi")

# dev version
devtools::install_git("https://gitlab-scm.partners.org/janickweberpals/smdi.git")
```

- Website (vignettes/articles): janickweberpals.gitlab-pages.partners.org/smdi
- Presentation slides: drugepi.gitlab-pages.partners.org/smdi-r-pharma-2023/smdi-r-pharma2023.html
- Presentation repository:
  - gitlab-scm.partners.org/drugepi/smdi-r-pharma-2023
  - github.com/janickweberpals/smdi-R-Pharma2023

Acknowledgments

Mass General Brigham

- Rishi J. Desai
- Robert J. Glynn
- Shamika More
- Luke Zabotka

Duke

- Sudha R. Raman
- Bradley G. Hammill

Kaiser WA

- Pamela A. Shaw

Harvard Pilgrim/SOC

- Darren Toh
- John G. Connolly Kimberly J.
- Dandreo Gegear

FDA

- Fang Tian
- Wei Liu
- Hana Lee
- Jie (Jenni) Li
- José J. Hernández-Muñoz

References

References cited in this presentation

1. Choi J, Dekkers OM, Cessie S le. A comparison of different methods to handle missing data in the context of propensity score analysis. *European Journal of Epidemiology*. 2019;34(1):23-36. doi:10.1007/s10654-018-0447-z
2. Mohan K, Pearl J. Graphical models for processing missing data. *Journal of the American Statistical Association*. 2021;116(534):1023-1037. doi:10.1080/01621459.2021.1874961
3. Tierney N, Cook D. Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations. 2023;105. doi:10.18637/jss.v105.i07
4. Buuren S van, Groothuis-Oudshoorn K. Mice: Multivariate imputation by chained equations in r. 2011;45:1-67. doi:10.18637/jss.v045.i03
5. Austin PC. Assessing covariate balance when using the generalized propensity score with quantitative or continuous exposures. *Statistical Methods in Medical Research*. 2018;28(5):1365-1377. doi:10.1177/0962280218756159
6. Hotelling H. The Generalization of Student's Ratio. *The Annals of Mathematical Statistics*. 1931;2(3):360-378. doi:10.1214/aoms/1177732979
7. Little RJA. A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*. 1988;83(404):1198-1202. doi:10.1080/01621459.1988.10478722
8. Sondhi A, Weberpals J, Yerram P, et al. A systematic approach towards missing lab data in electronic health records: A case study in non-small cell lung cancer and multiple myeloma. (accepted). *CPT Pharmacometrics Syst Pharmacol*.