COMP 551 – Applied Machine Learning
Lecture 24: Missing data and other loose ends

Instructor: Joelle Pineau (*jpineau@cs.mcgill.ca)*

Class web page: *www.cs.mcgill.ca/~jpineau/comp551*

Today: Missing data

- What’s missing?
  - Labels

    => Use unsupervised learning

  - A subset of observable features, in some of the data examples

    => Today’s lecture

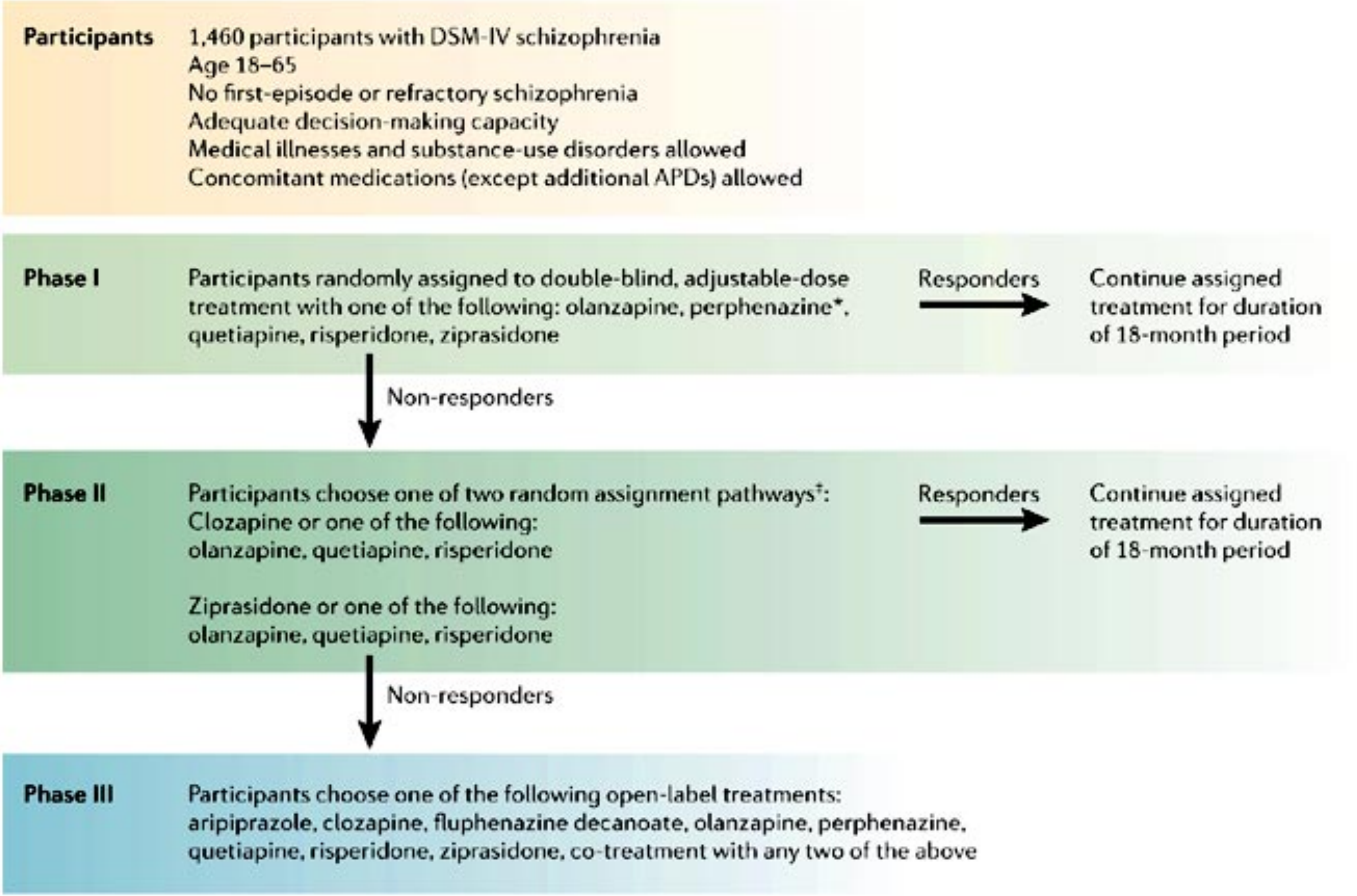
- Today’s lecture is not a comprehensive treatment of the topic, but rather a case study based on a recent research project:

  S.M. Shortreed, E. Laber, T.S. Stroup, J. Pineau, S.A. Murphy. "A multiple imputation strategy for sequential multiple assignment randomized trials". Statistics in Medicine. vol.33(24). pp.4202-4214. 2014.

A case study: CATIE study

- **CATIE** = Clinical Antipsychotic Trial of Intervention and Effectiveness.
  - 18 months, 1460 patients with schizophrenia.

- Data collected in a **Sequential Multiple Assignment Randomized Trial** (SMART).
  - Each patient is repeatedly randomized over time.
  - Each randomization occurs at a critical decision point (e.g. milestone in the disease process).
  - Timing and number of randomizations may vary across patients and depend on evolving patient-specific information.



Participants

1,460 participants with DSM-IV schizophrenia
Age 18–65
No first-episode or refractory schizophrenia
Adequate decision-making capacity
Medical illnesses and substance-use disorders allowed
Concomitant medications (except additional APDs) allowed

Phase I

Participants randomly assigned to double-blind, adjustable-dose treatment with one of the following: olanzapine, perphenazine*, quetiapine, risperidone, ziprasidone

Responders → Continue assigned treatment for duration of 18-month period

Non-responders ↓

Phase II

Participants choose one of two random assignment pathways[†]:
Clozapine or one of the following:
olanzapine, quetiapine, risperidone

Ziprasidone or one of the following:
olanzapine, quetiapine, risperidone

Responders → Continue assigned treatment for duration of 18-month period

Non-responders ↓

Phase III

Participants choose one of the following open-label treatments:
aripiprazole, clozapine, fluphenazine decanoate, olanzapine, perphenazine, quetiapine, risperidone, ziprasidone, co-treatment with any two of the above

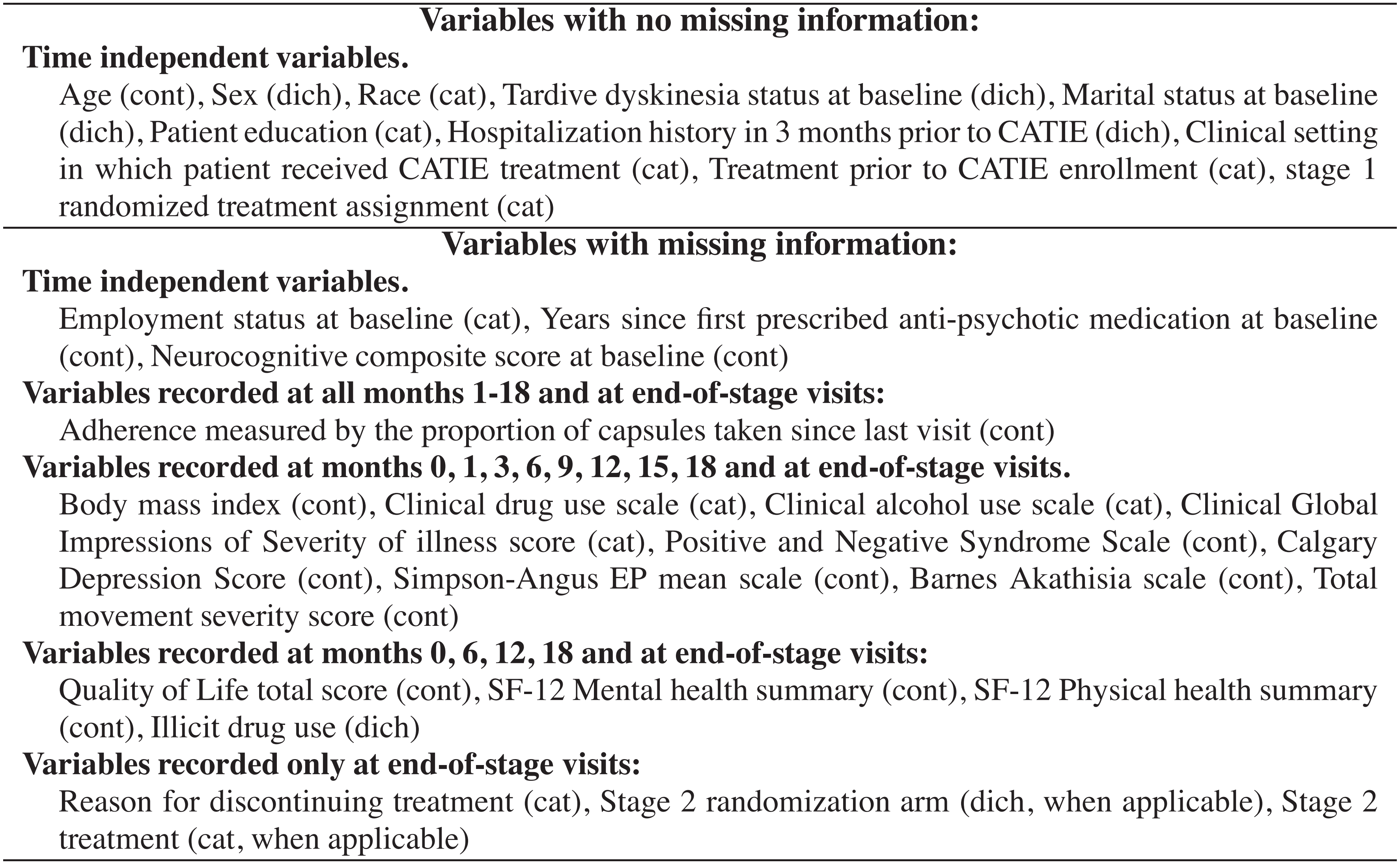
Performance measures

- Primary outcome:
  - Minimize “all-cause” treatment discontinuation (incl. efficacy, safety, tolerability).

- Secondary outcomes:
  - Symptoms, side effects, vocational and neurocognitive functioning, quality of life, caregiver burden, cost-effectiveness.

- **Scientific goal:** Find the sequence of treatments that produces the best performance according to these outcomes.

List of variables collected during CATIE



Variables with no missing information:

Time independent variables.

Age (cont), Sex (dich), Race (cat), Tardive dyskinesia status at baseline (dich), Marital status at baseline (dich), Patient education (cat), Hospitalization history in 3 months prior to CATIE (dich), Clinical setting in which patient received CATIE treatment (cat), Treatment prior to CATIE enrollment (cat), stage 1 randomized treatment assignment (cat)

Variables with missing information:

Time independent variables.

Employment status at baseline (cat), Years since first prescribed anti-psychotic medication at baseline (cont), Neurocognitive composite score at baseline (cont)

Variables recorded at all months 1-18 and at end-of-stage visits:

Adherence measured by the proportion of capsules taken since last visit (cont)

Variables recorded at months 0, 1, 3, 6, 9, 12, 15, 18 and at end-of-stage visits.

Body mass index (cont), Clinical drug use scale (cat), Clinical alcohol use scale (cat), Clinical Global Impressions of Severity of illness score (cat), Positive and Negative Syndrome Scale (cont), Calgary Depression Score (cont), Simpson-Angus EP mean scale (cont), Barnes Akathisia scale (cont), Total movement severity score (cont)

Variables recorded at months 0, 6, 12, 18 and at end-of-stage visits:

Quality of Life total score (cont), SF-12 Mental health summary (cont), SF-12 Physical health summary (cont), Illicit drug use (dich)

Variables recorded only at end-of-stage visits:

Reason for discontinuing treatment (cat), Stage 2 randomization arm (dich, when applicable), Stage 2 treatment (cat, when applicable)

Artificial CATIE dataset

G_0	W_0	P_0	$A_1 4$	W_1	P_1	C_1	A_2	P_2	W_2
Female	31.8	103	Perphenazine	23.4	77	SWITCHED	Ziprasidone	86	26.9
Male	29.4	108	Risperidone	18.2	102	STAYED	NA	88	19
Male	32.6	63	Olanzapine	35.2		STAYED	NA	85	38.2
Female		102	Quetiapine	34.6	99	SWITCHED	Olanzapine	77	
Female			Risperidone	20.8	96	SWITCHED	Olanzapine	71	31.6
Male	38.1	86	Perphenazine	28.7	75	STAYED	NA		
Female	31.1	80	Risperidone	22.8	89	SWITCHED	Clozapine		
Female	31.6	71	Olanzapine	21.1		STAYED	NA		
Male	25.1		Perphenazine	19.7	74	STAYED	NA		
Male	37.9	64	Olanzapine		36	STAYED	NA		
Female	28.7	91	Risperidone						
Male	37.8	65	Perphenazine						

W = Body-Mass Index
P = PANSS score (measure of symptom intensity)
A = Treatment assigned

Missing data in CATIE

- High study attrition: only 705 of 1460 stayed for full 18 months; 509 dropped out before entering stage 2.
  - High attrition is not unusual for studies of antipsychotics.
- Majority of missing data (78.1%) was due to attrition.

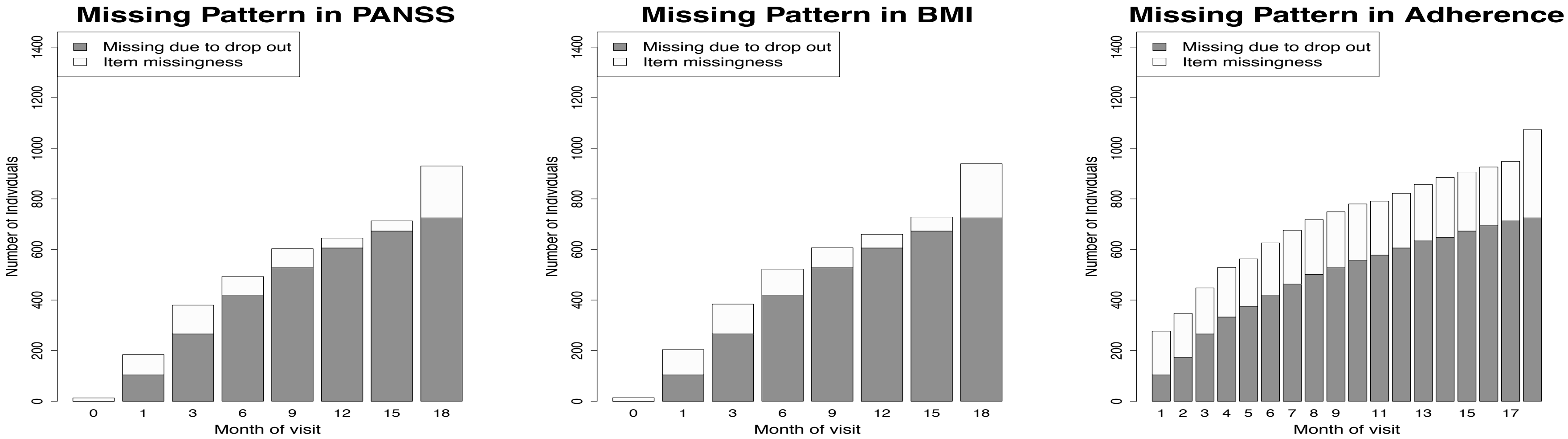
Missing data in CATIE

- High study attrition: only 705 of 1460 stayed for full 18 months; 509 dropped out before entering stage 2.
  - High attrition is not unusual for studies of antipsychotics.
- Majority of missing data (78.1%) was due to attrition.
- We observe a nearly monotone missing data pattern.
  - Monotone: missing data at time *t* -> missing data at time *t+1*.
- Distribution of most variables appears similar for participants that completed study and those that dropped out.

Missing data in CATIE

- Trend in the amount of missing data over time and proportion of missing data due to dropout are similar for many variables.

Figure 1. Bar plots showing the amount of missing data in the CATIE study. The total height of the bar displays the absolute number of people who have missing (a) PANSS, (b) BMI, and (c) adherence, as measured by pill count, at each of the monthly visits at which the scheduled variable was collected. The dark grey area represents individuals with missing values because they have dropped out of the study prior to that month. The unshaded area is the amount of item missingness in each variable.



(a) (b) (c)

Types of missing data

- Missing Completely at Random (MCAR)
  - A feature is missing at random, independent of the observed features or the output.

Types of missing data

- Missing Completely at Random (MCAR)
  - A feature is missing at random, independent of the observed features or the output.
- Missing at Random (MAR)
  - The missing value can depend on other observed variables, but not on the value of the missing feature itself.

Types of missing data

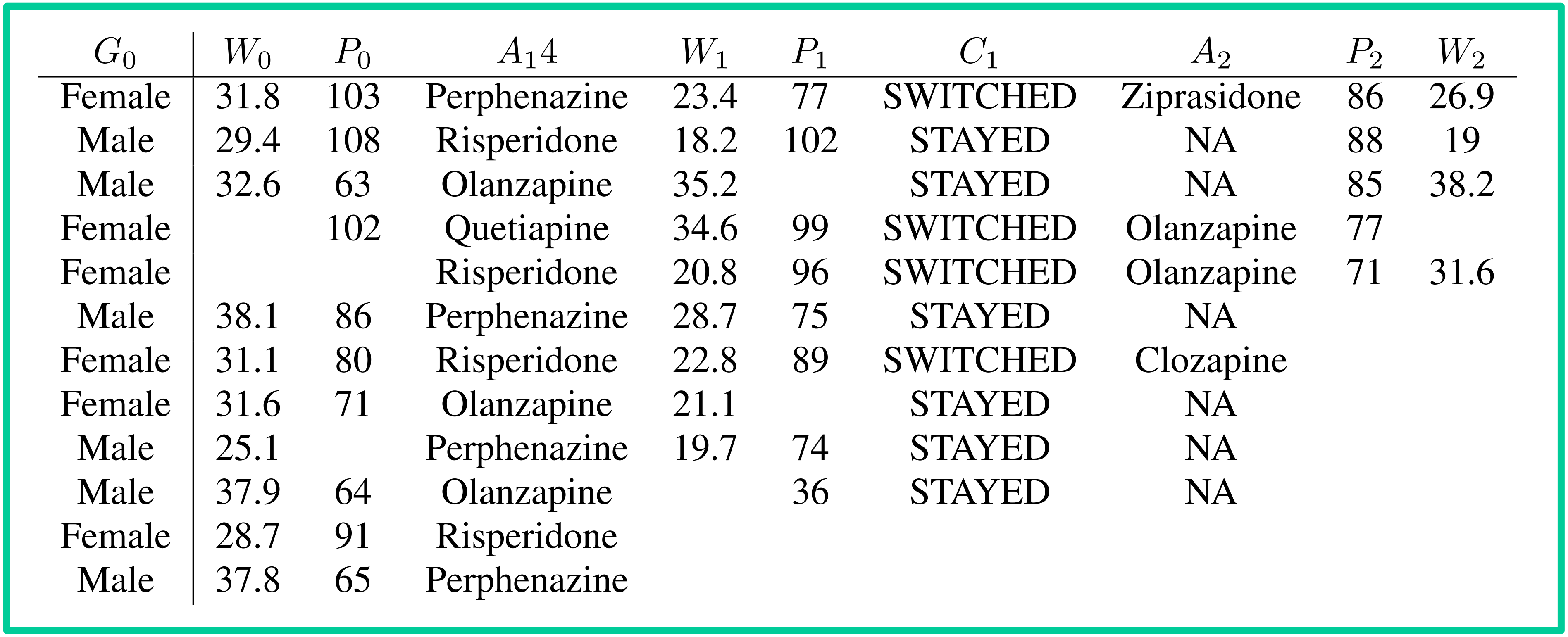
- Missing Completely at Random (MCAR)
  - A feature is missing at random, independent of the observed features or the output.
- Missing at Random (MAR)
  - The missing value can depend on other observed variables, but not on the value of the missing feature itself.
- Not Missing at Random (NMAR)
  - The missing value may depend on unobserved variables.

- In general: Hard to detect which case we are dealing with!

Strategies for missing data

Listwise deletion (Complete case analysis)

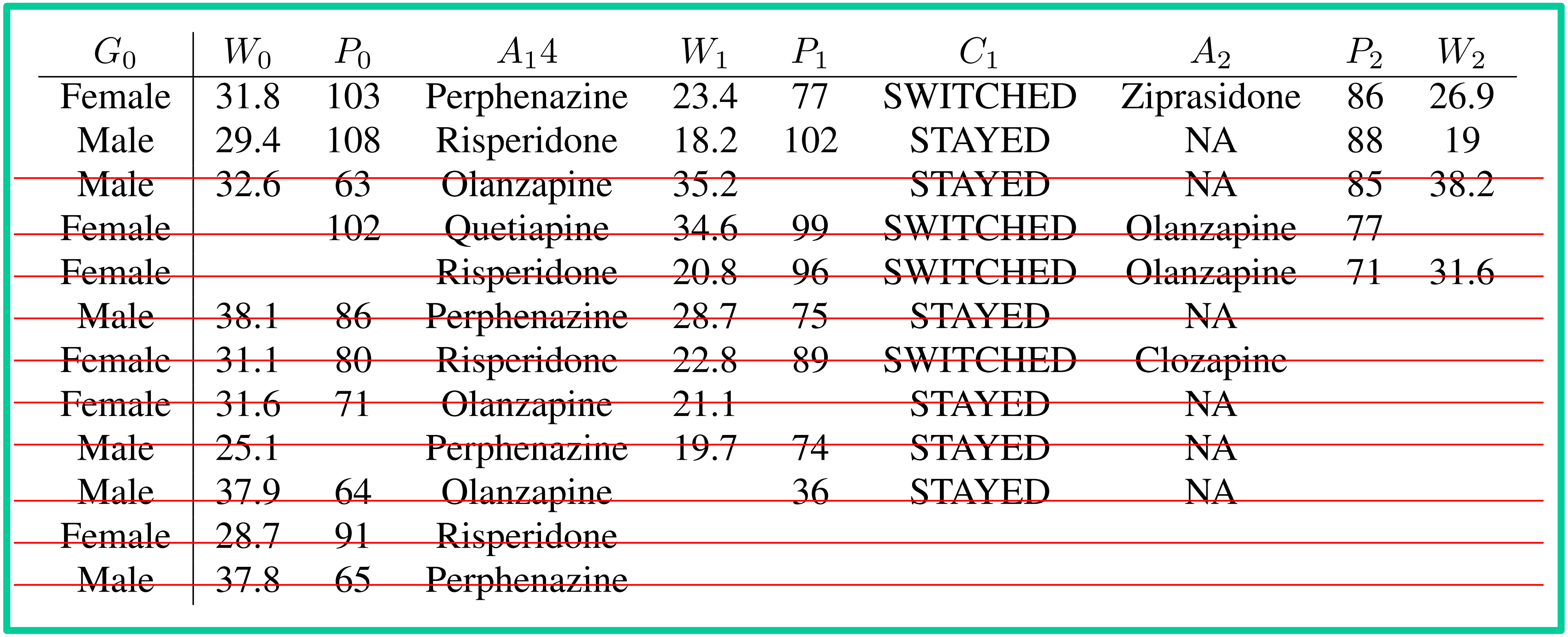
- Only use complete data points.
- Easy to implement!



G_0	W_0	P_0	A_14	W_1	P_1	C_1	A_2	P_2	W_2
Female	31.8	103	Perphenazine	23.4	77	SWITCHED	Ziprasidone	86	26.9
Male	29.4	108	Risperidone	18.2	102	STAYED	NA	88	19
Male	32.6	63	Olanzapine	35.2		STAYED	NA	85	38.2
Female		102	Quetiapine	34.6	99	SWITCHED	Olanzapine	77	
Female			Risperidone	20.8	96	SWITCHED	Olanzapine	71	31.6
Male	38.1	86	Perphenazine	28.7	75	STAYED	NA		
Female	31.1	80	Risperidone	22.8	89	SWITCHED	Clozapine		
Female	31.6	71	Olanzapine	21.1		STAYED	NA		
Male	25.1		Perphenazine	19.7	74	STAYED	NA		
Male	37.9	64	Olanzapine		36	STAYED	NA		
Female	28.7	91	Risperidone						
Male	37.8	65	Perphenazine						

Listwise deletion (Complete case analysis)

- Only use complete data points.
- Easy to implement!
- Wastes lots of data. Predictions may be biased if data is not MCAR.



G_0	W_0	P_0	A_14	W_1	P_1	C_1	A_2	P_2	W_2
Female	31.8	103	Perphenazine	23.4	77	SWITCHED	Ziprasidone	86	26.9
Male	29.4	108	Risperidone	18.2	102	STAYED	NA	88	19
~~Male~~	~~32.6~~	~~63~~	~~Olanzapine~~	~~35.2~~		~~STAYED~~	~~NA~~	~~85~~	~~38.2~~
~~Female~~		~~102~~	~~Quetiapine~~	~~34.6~~	~~99~~	~~SWITCHED~~	~~Olanzapine~~	~~77~~	
~~Female~~			~~Risperidone~~	~~20.8~~	~~96~~	~~SWITCHED~~	~~Olanzapine~~	~~71~~	~~31.6~~
~~Male~~	~~38.1~~	~~86~~	~~Perphenazine~~	~~28.7~~	~~75~~	~~STAYED~~	~~NA~~		
~~Female~~	~~31.1~~	~~80~~	~~Risperidone~~	~~22.8~~	~~89~~	~~SWITCHED~~	~~Clozapine~~		
~~Female~~	~~31.6~~	~~71~~	~~Olanzapine~~	~~21.1~~		~~STAYED~~	~~NA~~		
~~Male~~	~~25.1~~		~~Perphenazine~~	~~19.7~~	~~74~~	~~STAYED~~	~~NA~~		
~~Male~~	~~37.9~~	~~64~~	~~Olanzapine~~		~~36~~	~~STAYED~~	~~NA~~		
~~Female~~	~~28.7~~	~~91~~	~~Risperidone~~						
~~Male~~	~~37.8~~	~~65~~	~~Perphenazine~~						

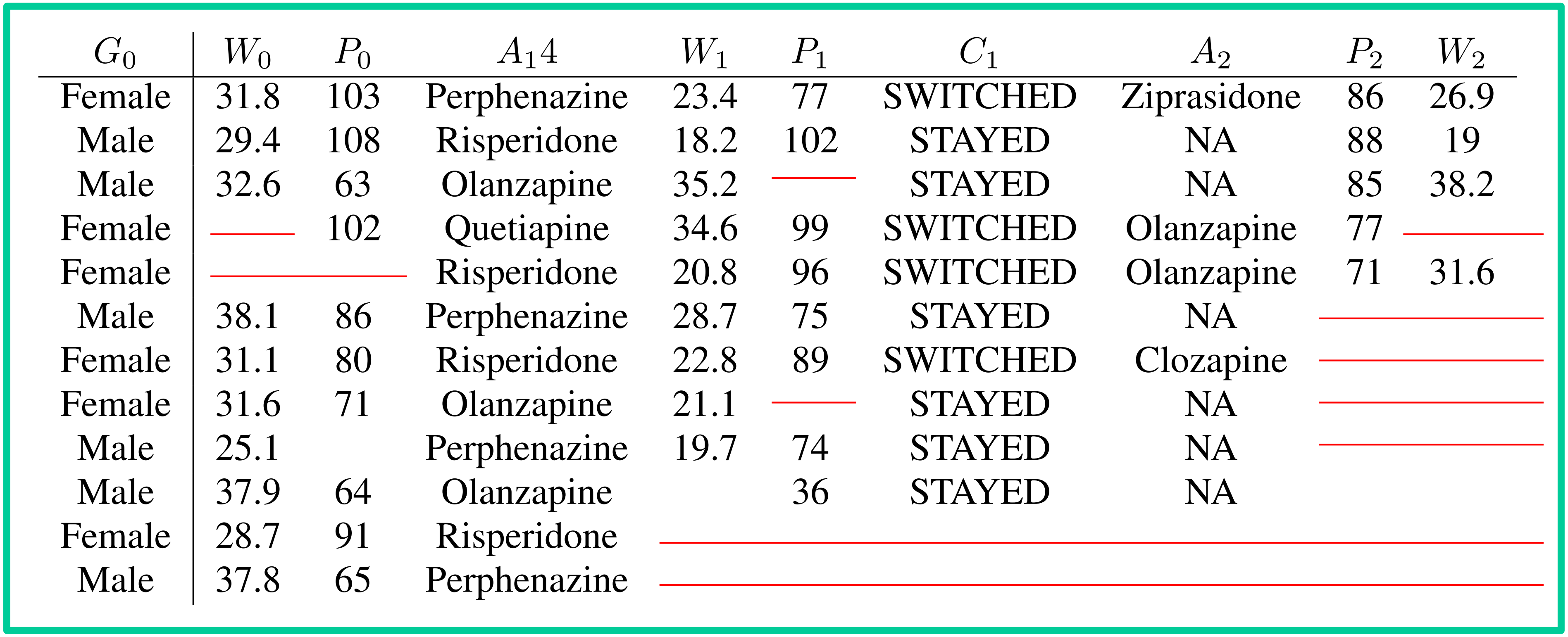
Pairwise deletion (Available case analysis)

- Use all cases in which the variables of interest are present.
    - E.g. Decision tree: evaluate test on x_i, using examples with that var.
- Uses as much information as possible.

G_0	W_0	P_0	$A_1 4$	W_1	P_1	C_1	A_2	P_2	W_2
Female	31.8	103	Perphenazine	23.4	77	SWITCHED	Ziprasidone	86	26.9
Male	29.4	108	Risperidone	18.2	102	STAYED	NA	88	19
Male	32.6	63	Olanzapine	35.2		STAYED	NA	85	38.2
Female		102	Quetiapine	34.6	99	SWITCHED	Olanzapine	77	
Female			Risperidone	20.8	96	SWITCHED	Olanzapine	71	31.6
Male	38.1	86	Perphenazine	28.7	75	STAYED	NA		
Female	31.1	80	Risperidone	22.8	89	SWITCHED	Clozapine		
Female	31.6	71	Olanzapine	21.1		STAYED	NA		
Male	25.1		Perphenazine	19.7	74	STAYED	NA		
Male	37.9	64	Olanzapine		36	STAYED	NA		
Female	28.7	91	Risperidone						
Male	37.8	65	Perphenazine						

Pairwise deletion (Available case analysis)

- Use all cases in which the variables of interest are present.
    - E.g. Decision tree: evaluate test on x_i, using examples with that var.
- Uses as much information as possible.
- Difficult to analyze since using different feature vectors. Bias if not MCAR.



G_0	W_0	P_0	A_14	W_1	P_1	C_1	A_2	P_2	W_2
Female	31.8	103	Perphenazine	23.4	77	SWITCHED	Ziprasidone	86	26.9
Male	29.4	108	Risperidone	18.2	102	STAYED	NA	88	19
Male	32.6	63	Olanzapine	35.2		STAYED	NA	85	38.2
Female		102	Quetiapine	34.6	99	SWITCHED	Olanzapine	77	
Female			Risperidone	20.8	96	SWITCHED	Olanzapine	71	31.6
Male	38.1	86	Perphenazine	28.7	75	STAYED	NA		
Female	31.1	80	Risperidone	22.8	89	SWITCHED	Clozapine		
Female	31.6	71	Olanzapine	21.1		STAYED	NA		
Male	25.1		Perphenazine	19.7	74	STAYED	NA		
Male	37.9	64	Olanzapine		36	STAYED	NA		
Female	28.7	91	Risperidone						
Male	37.8	65	Perphenazine						

Strategies for missing data

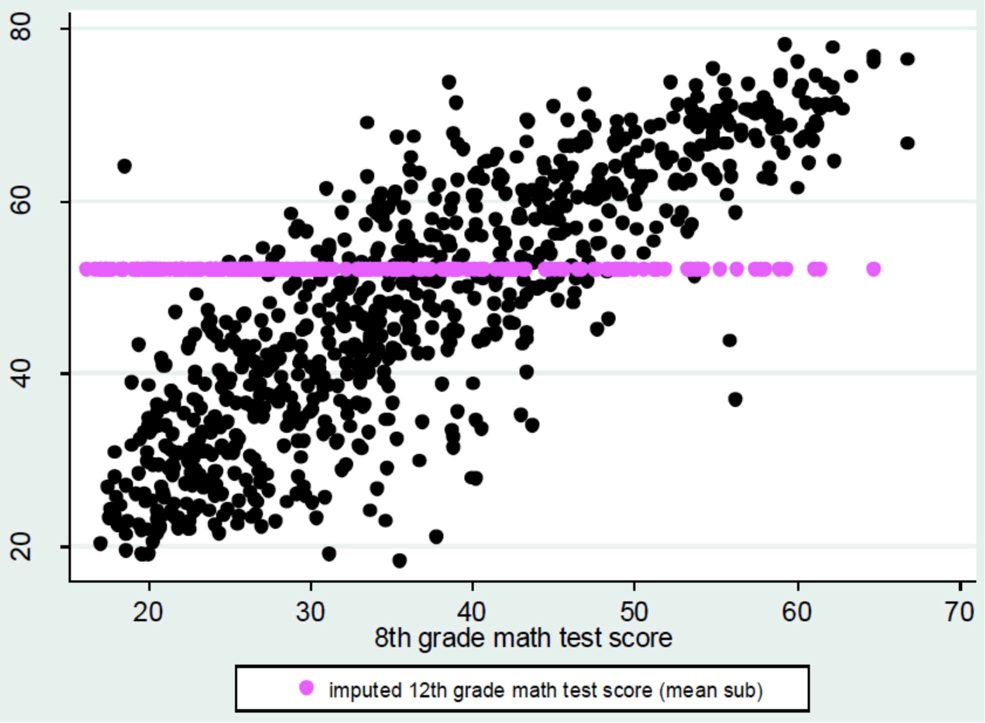
- **Deletion methods** => Remove cases (examples) from dataset
  - Listwise deletion
  - Pairwise deletion

- **Substitution methods** => Fill-in missing data

- **Model-based methods**

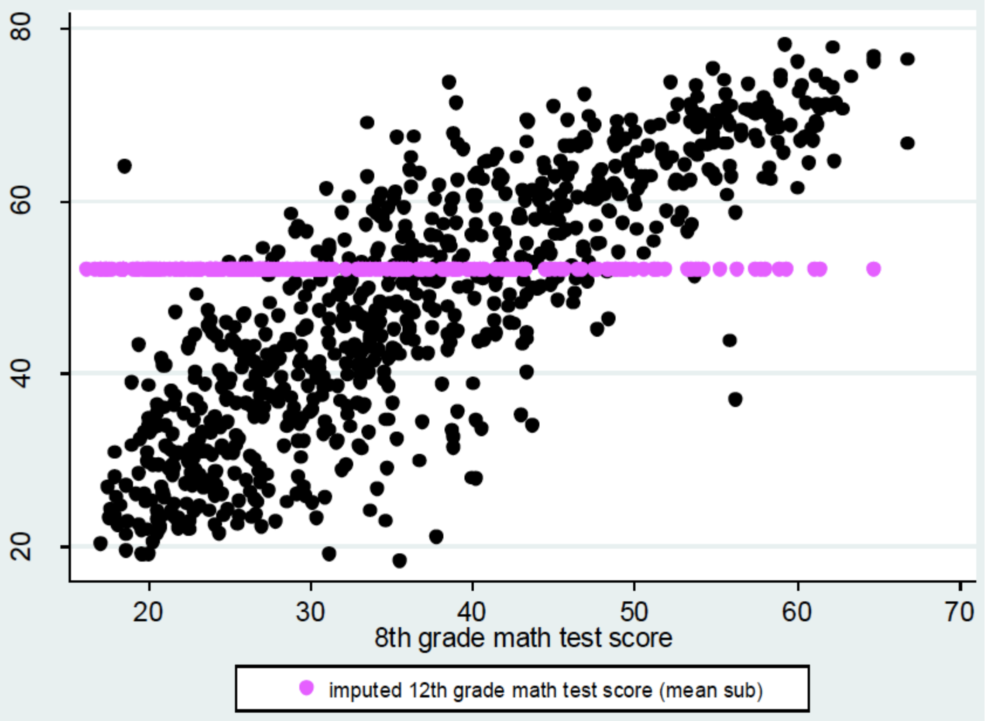
Mean / Mode substitution

- Replace missing value with sample mean or mode.
- Train learner as if all complete cases.



Mean / Mode substitution

- Replace missing value with sample mean or mode.
- Train learner as if all complete cases.



- Advantages:
  - Easy to implement!
- Disadvantages:
  - Bias unless MCAR.
  - Reduces variability.
  - Weakens covariance and correlation estimates in the data because it ignores relationship between variables.

Variable control

- **Add a binary indicator variable** (1 = value is missing; 0 = value is observed) to model missingness **for each variable**.
- Fill-in missing values using a constant (e.g. the sample mean).
- Train learner as in complete case, including indicator variables.

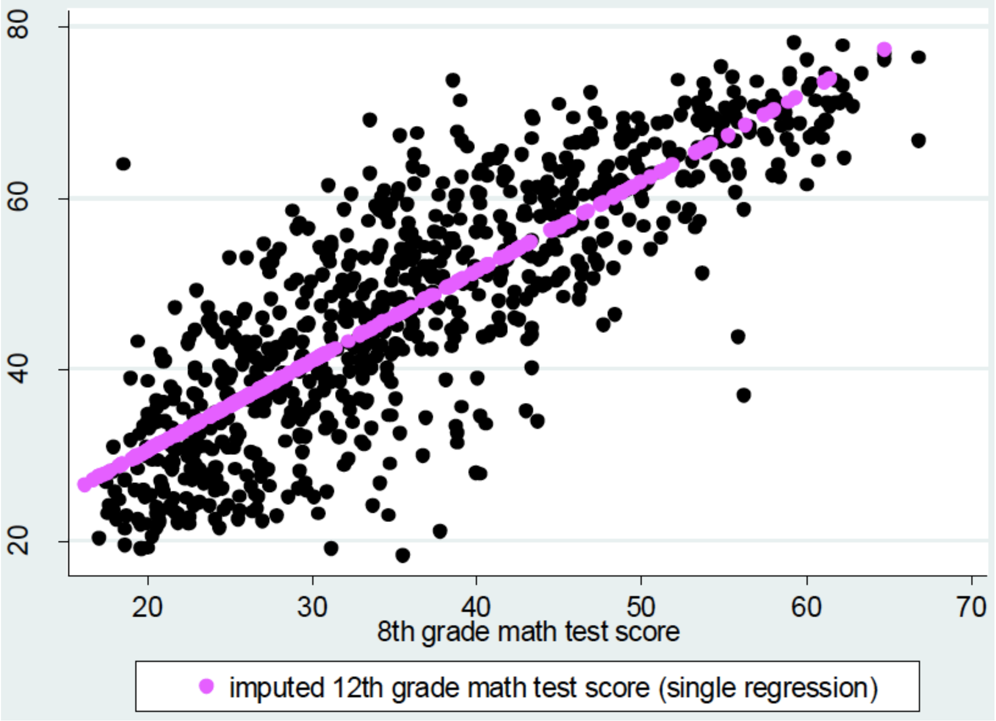
Variable control

- **Add a binary indicator variable** (1 = value is missing; 0 = value is observed) to model missingness **for each variable**.
- Fill-in missing values using a constant (e.g. the sample mean).
- Train learner as in complete case, including indicator variables.

- Advantage:
  - Uses all available information about missing observation.
- Disadvantage:
  - Results in biased estimates, unless MCAR.

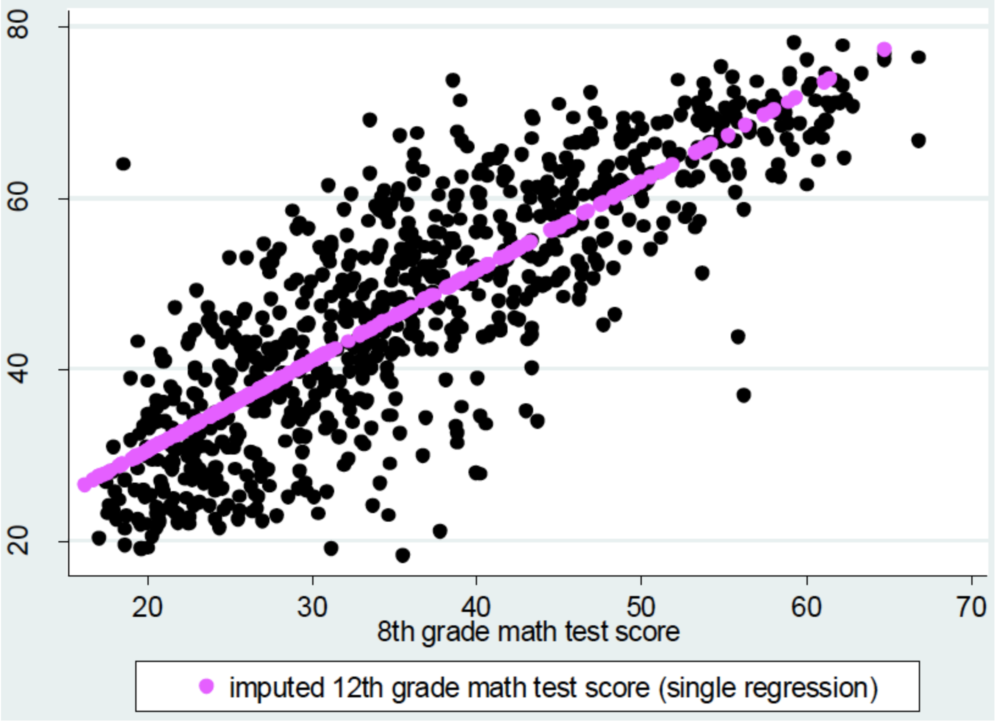
Regression imputation

- **Replace missing values with predicted value from a regression equation.**



Regression imputation

- **Replace missing values with predicted value from a regression equation.**



- Advantage:
  - Uses information from observed data.
- Disadvantage;
  - Overestimates model fit and correlation estimates. Weakens variance.

Strategies for missing data

- **Deletion methods** => Remove cases (examples) from dataset
  - Listwise deletion
  - Pairwise deletion
- **Substitution methods** (single imputation) => Fill-in missing data
  - Mean/mode substitution
  - Variable control
  - Regression imputation
- **Model-based methods** => Fill in missing data by building model of the data
  - Generative approach
  - Multiple imputation

Generative approach

- Assume a joint probabilistic model for the data.
- **Estimate the maximum likelihood setting for the missing data** using Expectation-Maximization.

- Advantages:
  - Uses full information to calculate likelihood.
  - Unbiased parameter estimation for MCAR/MAR cases.
- Disadvantages:
  - Converges to a local minima.

Multiple imputation

- Imputation: Data is “filled in” with predicted values from a trained regression model.
- Need a good regression model to get good imputations.

Multiple imputation

- Imputation: Data is “filled in” with predicted values from a trained regression model.
- Need a good regression model to get good imputations.

- Repeat imputation *k* times, producing *k* separate datasets
- Train predictor for each imputed (complete) dataset and merge results into one estimate (e.g. majority voting).

- This is the approach we implemented for CATIE.

Multiple imputation in CATIE

- Fit a (separate) conditional model for each variable.
- Algorithm:
  - Let $v_{t1}, \ldots v_{t,J}$ denote the variables collected at time t.
  - Order these variables according to amount of missingness.
  - Let $D_{t-1} = \{v_0, v_{1,1}, \ldots, v_{1,J1}, \ldots, v_{t-1,1}, \ldots, v_{t-1,Jt-1}\}$.

Multiple imputation in CATIE

- Fit a (separate) conditional model for each variable.
- Algorithm:
  - Let v_{t1}, … $v_{t,J}$ denote the variables collected at time t.
  - Order these variables according to amount of missingness.
  - Let $D_{t-1} = \{v_0, v_{1,1}, \ldots, v_{1,J1}, \ldots, v_{t-1,1}, \ldots, v_{t-1,Jt-1}\}$.
  - Estimate the joint posterior predictive distribution of the missing observations given the observed variables:

$$\int \cdots \int \prod_{t=1}^{T} \prod_{j=1}^{J_t} f(\mathbf{v}_{t,j}|\mathcal{D}_{t-1},\theta_{t,j})\pi(\theta_{t,j}|\mathcal{D}_{t-1},\mathbf{v}_{t,j,\mathrm{obs}},\theta_{1,1},\ldots,\theta_{t-1,J_t},\ldots,\theta_{t,j-1})d\theta_{t,j}$$

  - First term is the conditional on the current variable. Second term is the prior on the parameters of the distribution (θ).
  - The posterior is estimated by sampling, time step by time step

Multiple imputation in CATIE

- Using separate models for each variable is computationally advantageous (compared to full joint distribution over all variables.)
- But can lead to unrealistic fluctuations in some variables over time.

Multiple imputation in CATIE

- Using separate models for each variable is computationally advantageous (compared to full joint distribution over all variables.)
- But can lead to unrealistic fluctuations in some variables over time.

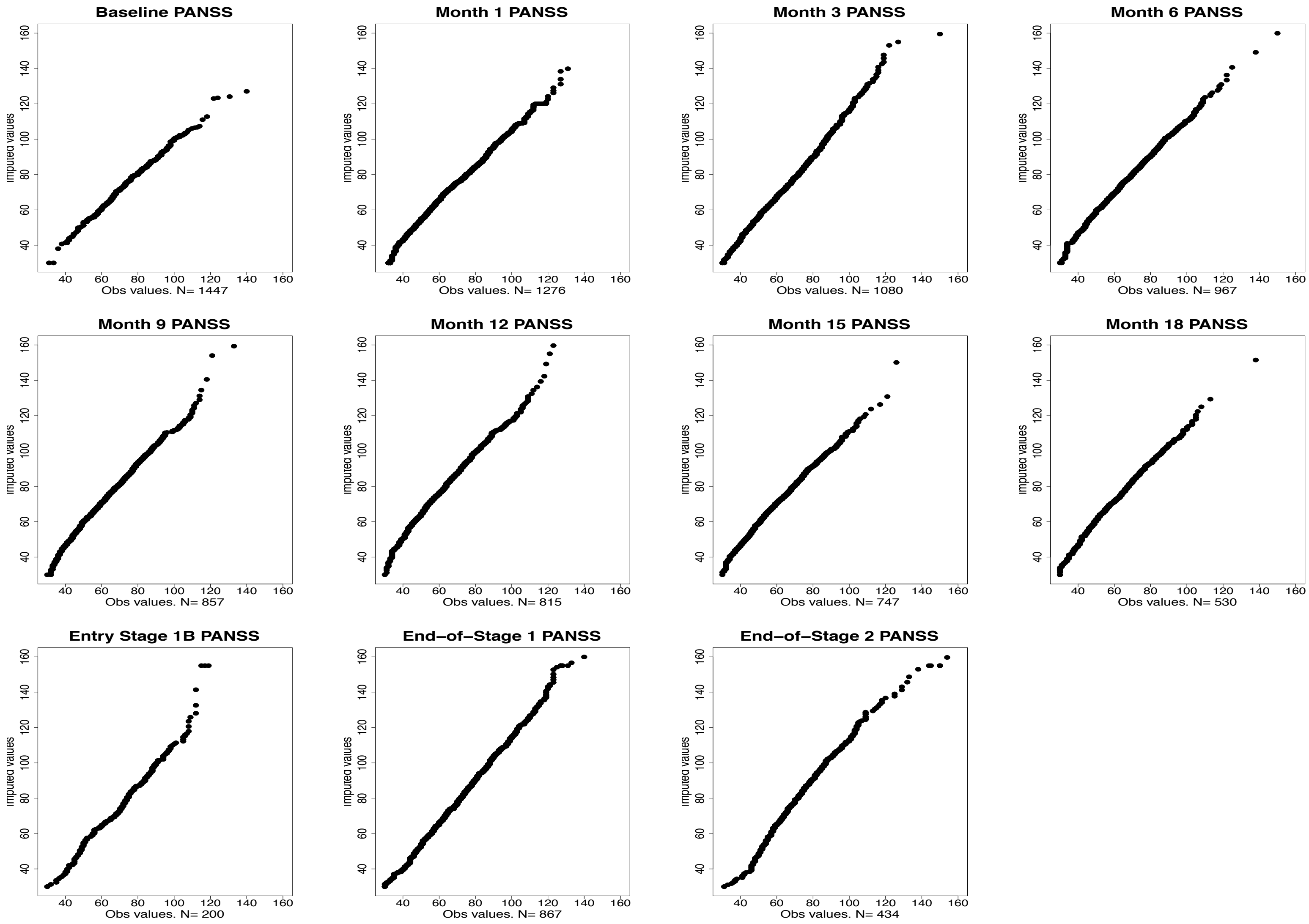
- Challenge: impose smoothness constraint (over time) on some variables.
- Solution: Use spline regression to enforce smoothness over time on the conditional mean.

Multiple imputation in CATIE

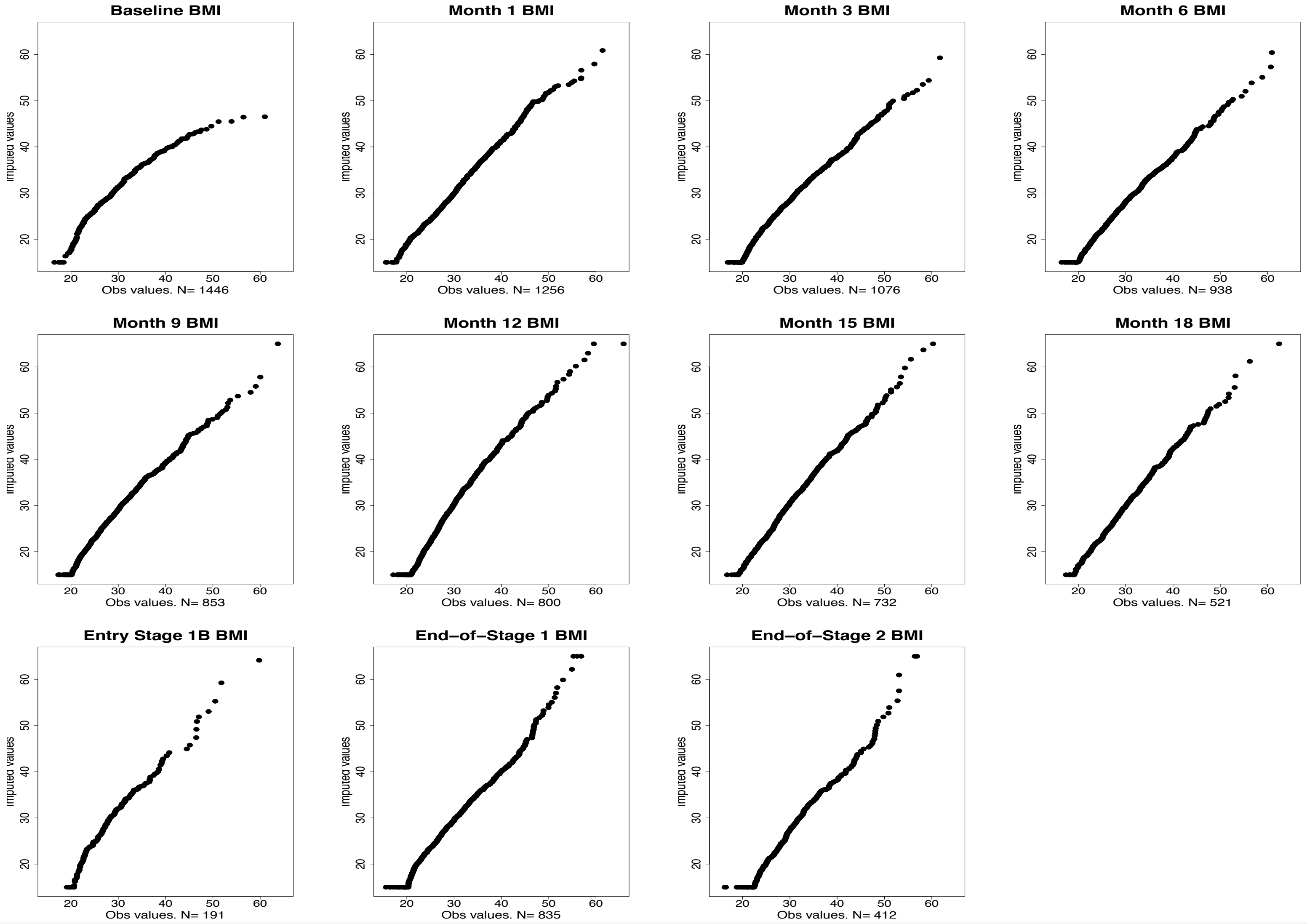
Overall imputation strategy:

1. Impute baseline variables (only 3% of data is missing).
2. Impute stage transition times. Use single imputation for this.
3. Impute end-of-stage variables.
   - Pool data over multiple time-windows (months) to get better estimates.
4. Impute randomly assigned treatment (especially for stage 2).
5. Impute additional missing time-varying information.

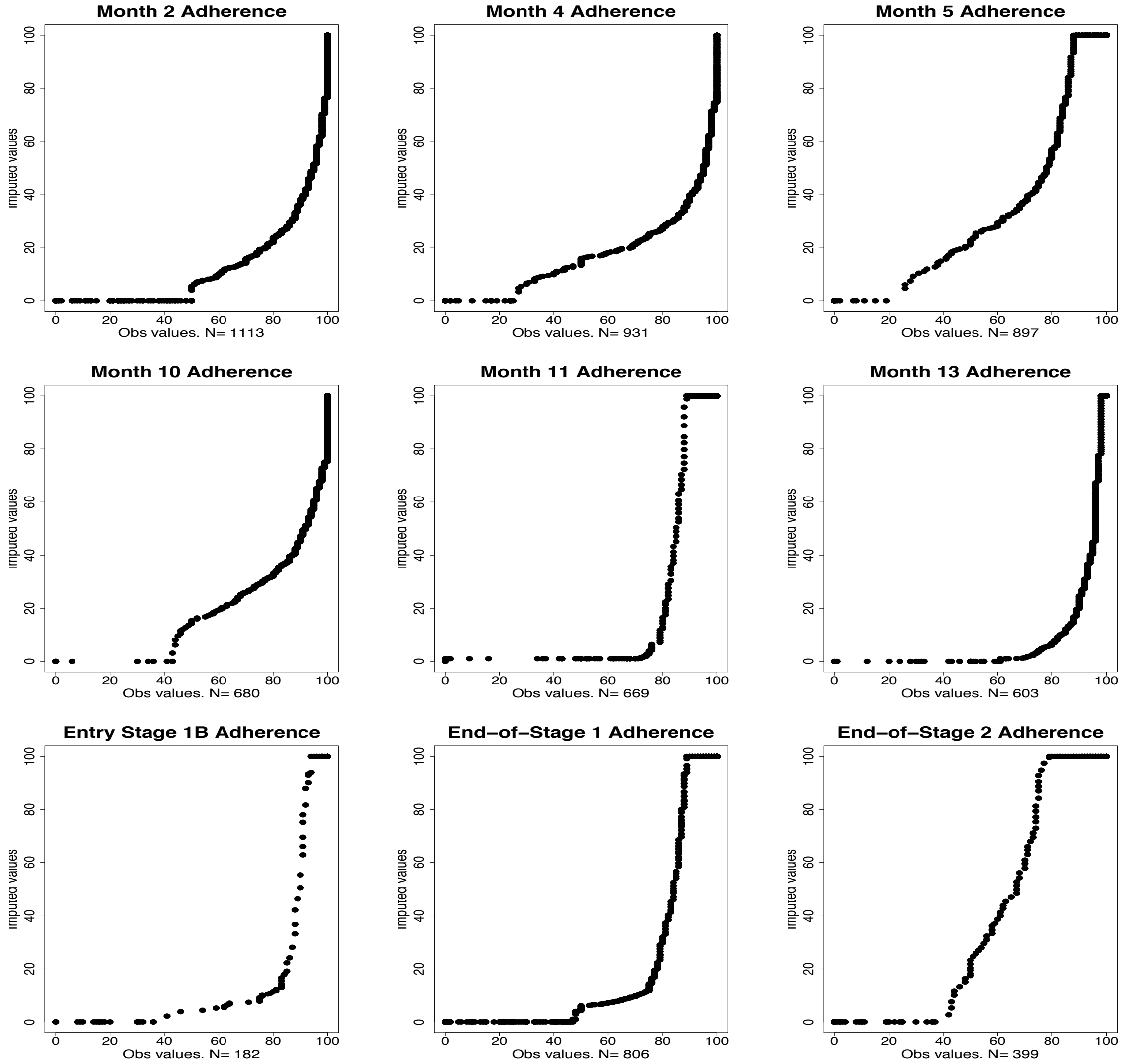
Imputed vs Observed PANSS scores



Imputed vs Observed BMI values

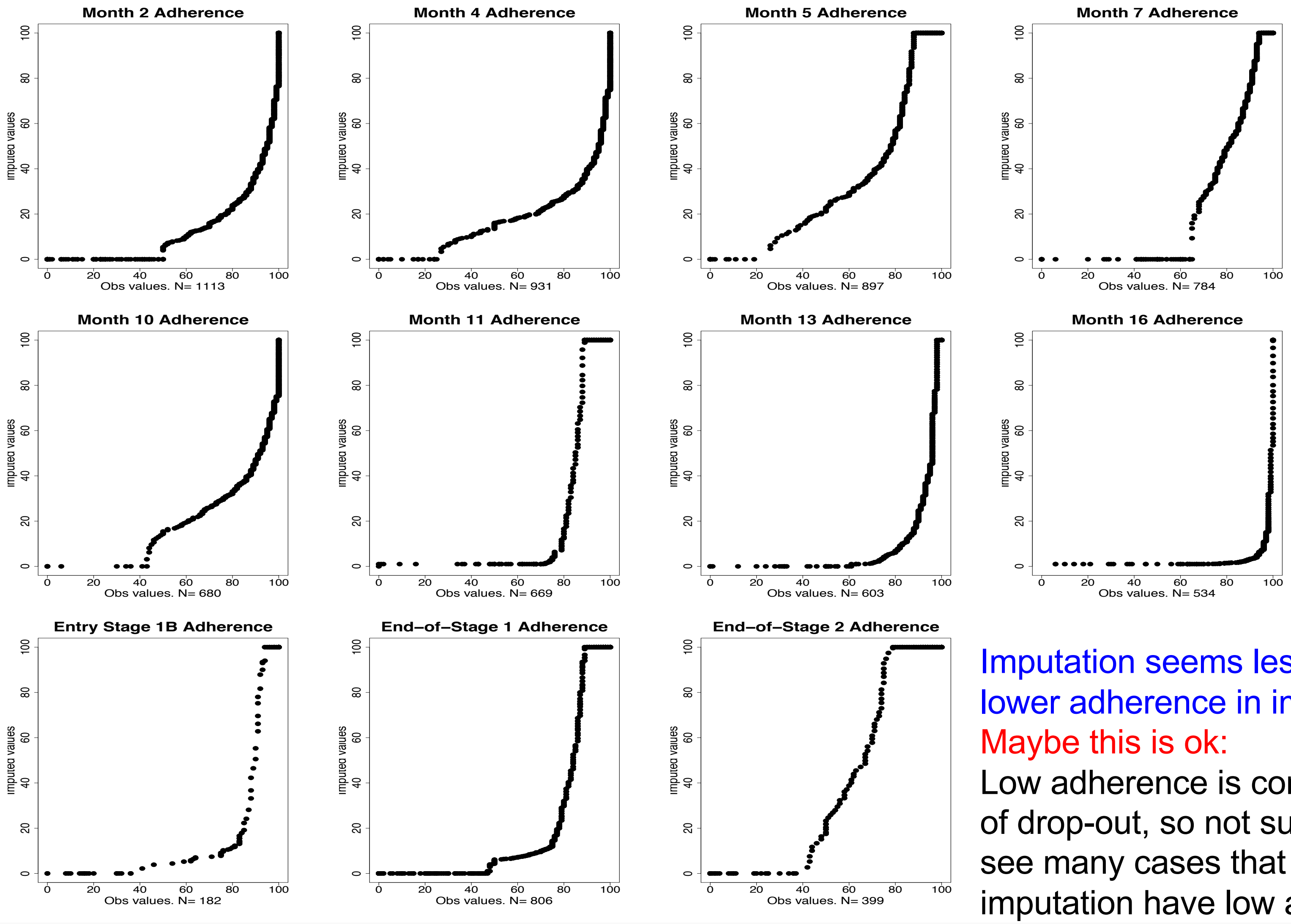


Imputed vs Observed Adherence



Imputation seems less accurate: lower adherence in imputed data.

Imputed vs Observed Adherence



Imputation seems less accurate: lower adherence in imputed data.

Maybe this is ok:

Low adherence is common cause of drop-out, so not surprising to see many cases that needed imputation have low adherence.

CATIE analysis with imputed data

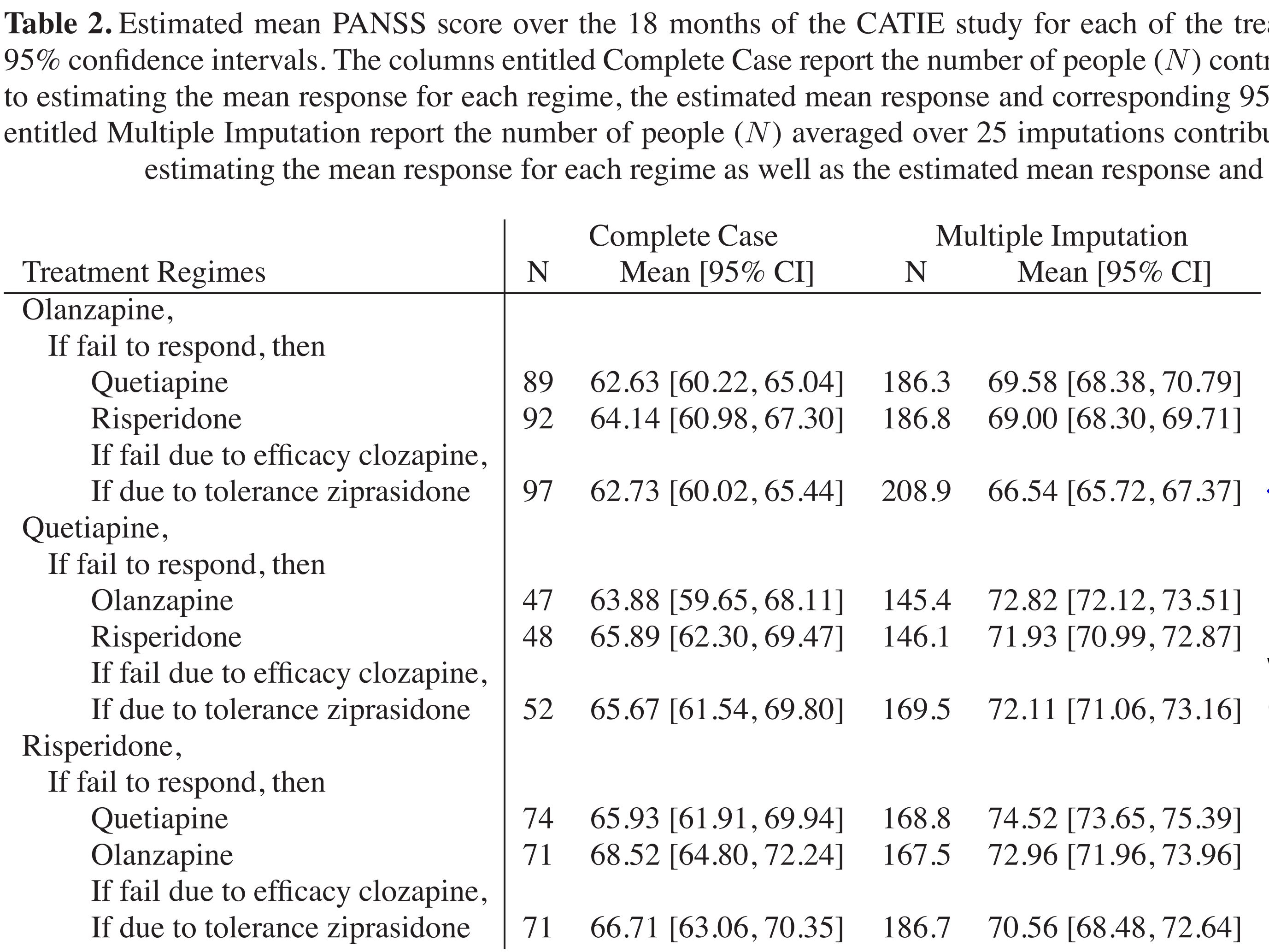


Table 2. Estimated mean PANSS score over the 18 months of the CATIE study for each of the treatment regimes and 95% confidence intervals. The columns entitled Complete Case report the number of people (N) contributing information to estimating the mean response for each regime, the estimated mean response and corresponding 95% CI. The columns entitled Multiple Imputation report the number of people (N) averaged over 25 imputations contributing information to estimating the mean response for each regime as well as the estimated mean response and 95% CI.

	Complete Case		Multiple Imputation	
Treatment Regimes	N	Mean [95% CI]	N	Mean [95% CI]
Olanzapine,				
If fail to respond, then				
Quetiapine	89	62.63 [60.22, 65.04]	186.3	69.58 [68.38, 70.79]
Risperidone	92	64.14 [60.98, 67.30]	186.8	69.00 [68.30, 69.71]
If fail due to efficacy clozapine,				
If due to tolerance ziprasidone	97	62.73 [60.02, 65.44]	208.9	66.54 [65.72, 67.37]
Quetiapine,				
If fail to respond, then				
Olanzapine	47	63.88 [59.65, 68.11]	145.4	72.82 [72.12, 73.51]
Risperidone	48	65.89 [62.30, 69.47]	146.1	71.93 [70.99, 72.87]
If fail due to efficacy clozapine,				
If due to tolerance ziprasidone	52	65.67 [61.54, 69.80]	169.5	72.11 [71.06, 73.16]
Risperidone,				
If fail to respond, then				
Quetiapine	74	65.93 [61.91, 69.94]	168.8	74.52 [73.65, 75.39]
Olanzapine	71	68.52 [64.80, 72.24]	167.5	72.96 [71.96, 73.96]
If fail due to efficacy clozapine,				
If due to tolerance ziprasidone	71	66.71 [63.06, 70.35]	186.7	70.56 [68.48, 72.64]

← Significantly better PANNS score.

No significant results with the Complete Case analysis.

Final comments

- Missing data can cause significant bias in analysis.
- Many methods for handling missing data; in general, need to understand your data and missingness pattern to figure out what technique is appropriate.
- EM algorithm can be used to estimate parameters of generative model and fill-in missing data.
- Multiple imputation is a successful method for cases with structural missingness, but requires significant modeling effort.

Other courses in machine learning

- **COMP-550(?): Natural language processing**
- **COMP-553: Game theory**
- **COMP-652: (Advanced) Machine learning**
  - Active learning, learning theory, graphical models, time-series.
- **COMP-767: Reinforcement learning**
  - Reinforcement learning theory, algorithms and applications.
- **ECSE-626: Statistical computer vision**
  - Probabilistic models and learning algorithms for computer vision.

- **IFT 6266 (@UdeM): Algorithmes d'apprentissage**
- **IFT 6085 (@UdeM): Advanced Structured Prediction**

Final project guidelines

- Report should contain:
  - Abstract (1 paragraph)
  - Introduction (1/2 page)
  - Technical summary of the paper selected (1/2 page)
  - Reproducibility methodology: what you reproduce, why, how (1-2 page).
  - Empirical results (with tables/graphs) (1-2 pages).
  - Discussion: see Reproducibility metrics in Lecture 23, slide 32 (1/2-1 page).
  - Conclusions of your analysis, limitations of your approach, open questions, suggestions for additional work (1 paragraph).
  - Append your Open Review (~1 page)
- Presentation: Summarize key points from above. Should have defined reproducibility methodology. Not expected to be done results. Max 4-5 slides.
- Open Review:
  - Post an executive summary (~1 page) of your report, you can include link to your full report and code (e.g. github repo).

Final notes

- **Project #3**
  - Peer reviews due on Thursday (I think – check CMT).
- **Project #4:**
  - Don’t forget to sign up for the challenge!

    https://docs.google.com/forms/d/1GAZnZWYW2suf6Z9polBlTQvTvMJIjkMy7CNyMapNKuY/edit?ts =59d53577
  - Pick a presentation slot; so far 21 teams signed up.

    https://docs.google.com/spreadsheets/d/1G_wGgR7leHvfr2TSri_IrMVZwXZGXgtx-nlik-4GSZo/edit#gid=0
  - Submit your slides for the presentation:

    https://drive.google.com/drive/folders/15AtV4cjE2ZIj5KgzG4vDm8QLkcN720Mp?usp=sharing
  - Final submission Dec.15 on CMT (report&code) and OpenReview (review).
- **Midterm:** Grades will be posted on MyCourses soon; available for viewing during office hours. Times will be posted on discussion board.
- **Quizzes**: Max 1pt per quiz. Max 5pts total (=5%), from the 12 quizzes.
- **Course evaluations** now available on Minerva. Please fill out!