Missing Values & Statistical Modelling

Krishnendu Biswas
Krishnendu.biswas@cognizant.com
+91 98196 90581

03rd December 2016
Mumbai

Cognizant

Disclaimer:

Views reflected are personal and DO NOT reflect the views of Cognizant Technology Solutions

Acknowledgement:

Thanks my colleagues **Neha Singh** and **Pankaj Tiwari**.

Cognizant

Flow:

- Problem statement
- Feasible solutions
- A Case Study

Lets start with an example:

- A randomized, double-blind study to compare nifedipine-GITS and verapamil-SR on hemodynamics, left ventricular mass, and coronary vasodilatory in patients with advanced hypertension .
- Fifty-four patients were randomized after the placebo run-in phase. 'Twenty-four failed to complete the (six-month) trial, and thus were not included for analysis because of
    1. withdrawal for symptomatic adverse effects,
    2. lack of response, and
    3. poor compliance

'Consequently, there were 30 subjects with sufficient data sets for inclusion in analyses.'

Lets start with an example:

- Result:
  - 24 patients out of the 30 completers achieved the desired endpoint
  - 80% were reported with effective control of BP and no side effects
- Is it really intended to estimate the chance for patients to have effective control of BP with no side effects ?
  - Shouldn't the denominator be 54 ?
- The true answer should be $(24^{(r)}+?)/(30+24^{(d)}) = (24+?)/54$

Lets start with an example:

- Options:
  - (a) none of the dropouts would respond
  - (b) all of the dropouts would respond
  - (c) An estimate between the extremes is (c): to substitute the unknown number by $24^{(d)} \times (24^{(r)}/30) = 24^{(d)} \times 0.80 = 19.2$, where $24^{(r)}/30 = 80\%$ is the proportion of responders among those who completed the trial.
- Unsurprisingly, when we do the calculation, the estimate becomes $(24^{(r)}+19.2)/54 = 43.2/54 = 80\%$, the same answer as that using only the completers. In fact, a simple algebra can show that this is always so.
- We should not be confused with these two '80%'. The former 80% is a wrong summary number; the latter is an estimate of the quantity of interest with a particular assumption about the missing data.

Lets start with an example:

- Thoughts:
  - It is important to record and report the reasons for withdrawal and the number of subjects in each category of withdrawal according to their treatment group
  - The reasons for patients dropping out can be used to help properly assess the nature of the missing data.
  - For example, if all the dropouts were because of a lack of response or side effects, then the calculation in (a) would be appropriate. In statistical terms, they would be called informative missing data. This is because useful information can be found in the reason for the dropout and this can be used to estimate the true response

Effects of Missing Data

- Reduction of data=>Reduction of power
- Loss of non-completers=> underestimation of variability
- Bias – why are the data missing? i.e. is the reason the data are missing related to treatment? If so, could lead to biased estimate of trt effect. Usually have to be conservative.

Handling of Missing Data

- Plan/try to **avoid** missing data by considering the following:
  - Length of study
  - Number of visits
  - Better follow-up (retrieving drop outs)
  - Amount of data being collected
  - Incentives for investigators to complete all visits for subjects
- Pre-specify in protocol how missing data will be dealt with (especially for primary analysis).
  - “Completed Case Analysis” – analyse only the subjects with complete data. Not recommended for primary analysis in a confirmatory trial
  - “Imputation” – LOCF, best & worst case analyses, use mixed effects models

Classification of Missing Data

- Missing Completely At Random (MCAR)
  - The reason for data being missing does not depend on the observed or the unobserved missing data

- Missing At Random (MAR)
  - The reason for data being missing may depend on observed data but not on the unobserved missing data
  - MCAR implies MAR but not the other way round

- Missing Not At Random (MNAR)
  - The reason for data being missing depends on the unobserved missing data (or on observed data not taken into account in the model)

Examples of Missing Data

Clinical trial randomly assigned 100 patients with major depression to an experimental drug (D) or to placebo (P).

Participants completed the Hamilton depression rating scale (HAMD) at baseline and again after 9-week treatment.

Suppose that some final surveys were missing—not completed.

- Example: a participant flips a coin to decide whether to complete the depression survey (MCAR)
- Example: male participants are more likely to refuse to fill out the depression survey, but it does not depend on the level of their depression. (MAR)
- Example: participants with severe depression, or side-effects from the medication, were more likely to be missing at end. (MNAR)

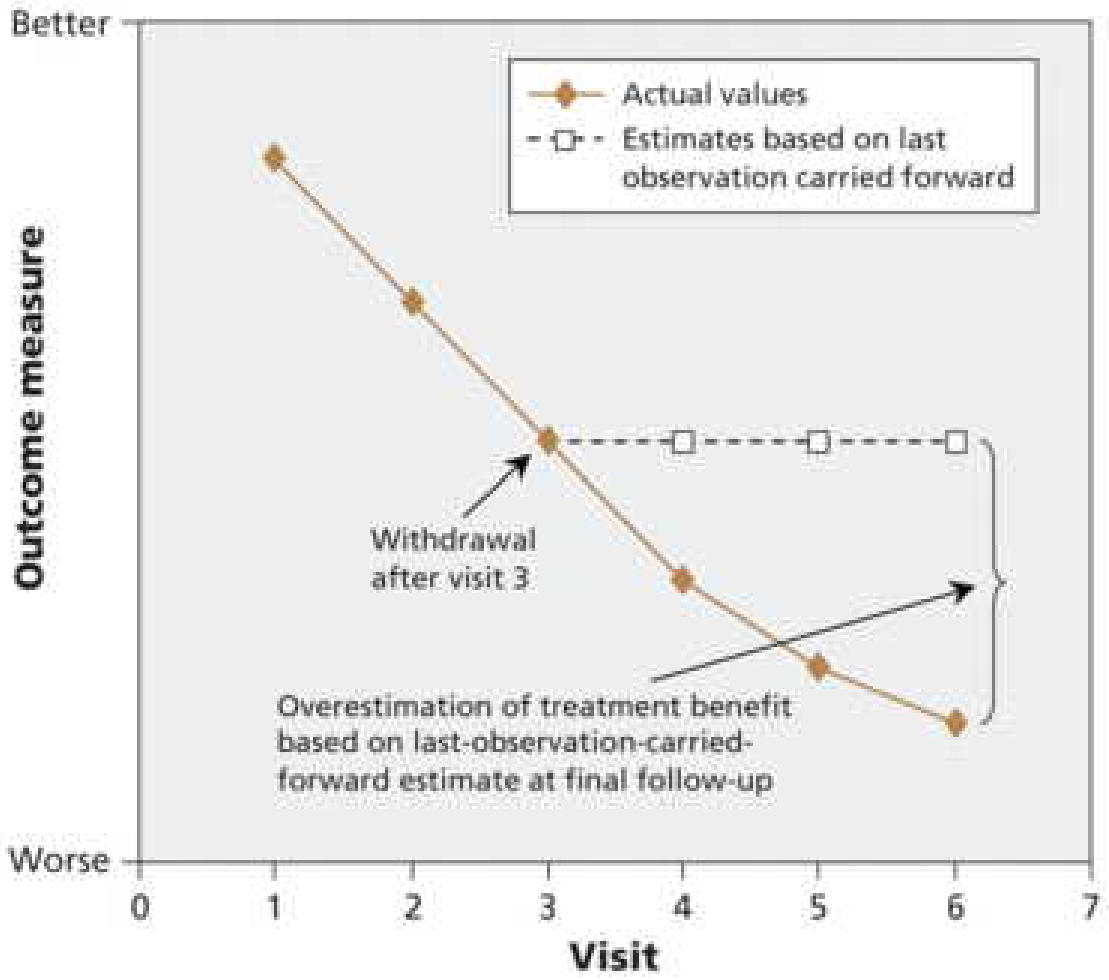
Limitations of some Traditional Methods

List-wise Deletion / Complete case analysis	• If 5 individuals have missing scores on one or more variables, we simply omit those individuals from the analysis. • Deletion often results in a substantial decrease in the sample size available • Bias ((For example when low income individuals are less likely to report their income level, the resulting mean is biased in favor of higher incomes.)
Pairwise Deletion	• If one participant reports his income and life satisfaction index, but not his age, he is included in the correlation of income and life satisfaction, but not in the correlations involving age. • The problem with this approach is that the parameters of the model will be based on different sets of data, with different sample sizes and different standard errors • It has been suggested that if there are only a few missing observations it doesn't hurt anything to use pairwise deletion.
LOCF	• LOCF ignores trajectory - whether the subject was improving or getting worse at the time of dropout. • LOCF freezes outcome values at the last observation, thereby creating an apparent stabilization of disease, symptoms and/or function in dropouts (problematic in progressive conditions)

Limitations of some Traditional Methods

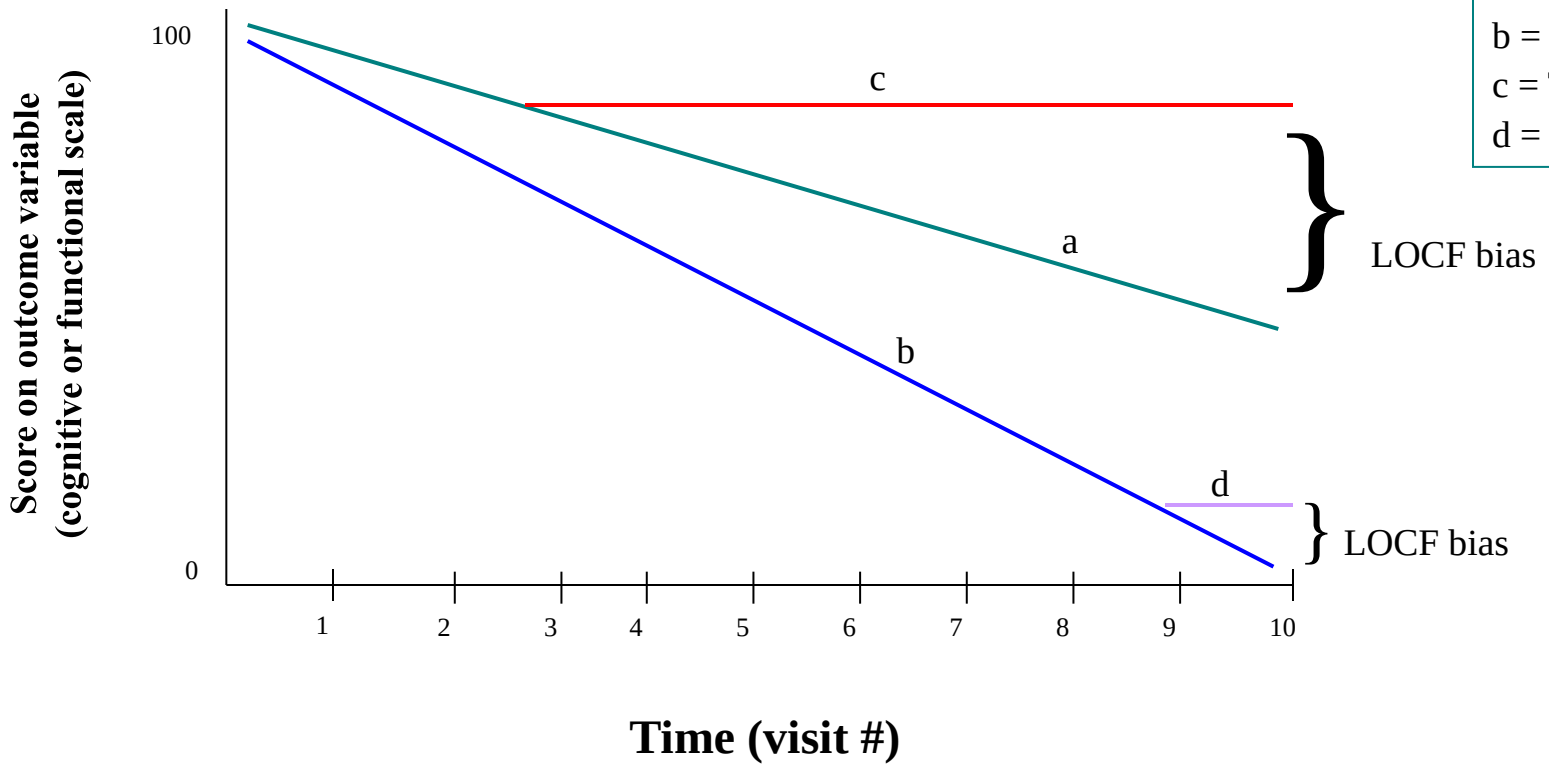
Mean Imputation	• It adds no new information. • In addition, such a process leads to an underestimate of error
Regression Substitution	• Better than mean imputation. Atleast imputed value is in some way conditional on other information we have about the person. • By substituting a value that is perfectly predictable from other variables, we have not really added more information but we have increased the sample size and reduced the standard error.

Figure : Potential impact of last-observation-carried-forward analysis in longitudinal randomized controlled trials in chronic progressive diseases



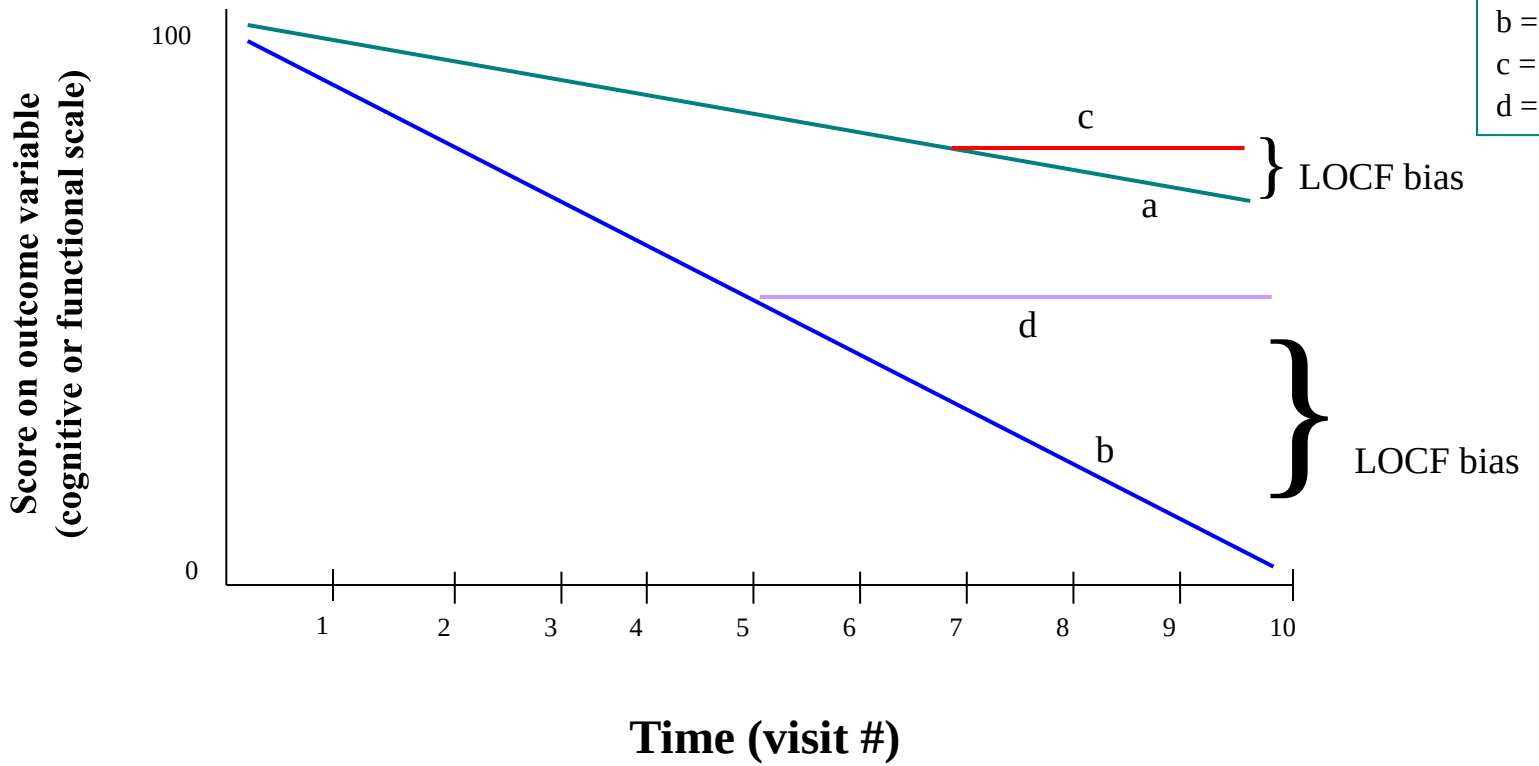
Differential LOCF bias if have greater or earlier dropouts in *treatment group* than control group
(Effect measured by LOCF [c – d] > True effect [a – b] resulting in exaggerated positive effect - biased in favor of treatment)

a = True treatment values
b = True control values
c = Treatment values altered by LOCF
d = Control values altered by LOCF



Differential LOCF bias if have greater or earlier dropouts in *control group* than treatment group
(Effect measured by LOCF [c – d] < True effect [a – b] resulting in underestimate of effectiveness - biased against treatment)

a = True treatment values
b = True control values
c = Treatment values altered by LOCF
d = Control values altered by LOCF



Multiple Imputation (MI) Method

- A simulation based approach where missing values are replaced by multiple Bayesian draws from the conditional distribution of missing data given the observed data and covariates, creating multiple completed data sets
- In multiple imputation we generate imputed values on the basis of existing data
- Two random processes involved
  - First, the imputed value contains a random component from a standard normal distribution
  - Second, the parameter estimates used in imputing data are a random draw from a posterior probability distribution of the parameters.

CASE STUDY

Primary Efficacy Endpoint Details

- Endpoint: PSSI 90 response, based on PSSI score
- Type: Response in Yes, No
- Analysis Timepoint : 12 week
- Responder : subject achieving ≥90% improvement (reduction) in PSSI score compared to baseline
- Non Responder : subject not achieving ≥90% improvement (reduction) in PSSI score compared to baseline
- Primary analysis method :
  - CMH test
  - 95% CI, for the difference between two treatments in the proportion of PSSI 90 responders, was calculated using the normal approximation to the binomial distribution.

[PSSI:Psoriasis Scalp Severity Index (**PSSI**)]

Analysis

- Primary Efficacy Endpoint Details
  - Comparison Between : Test drug and Placebo
  - Adjustment factor : Body weight (<90 kg, ≥90 kg)
  - In the primary analysis PSSI 90 response (yes, no) at Week 12, a subject with a missing assessment was considered as a non-responder (no).

- Sensitivity Analysis for Primary Efficacy Endpoint
  - Sensitivity analysis of primary efficacy endpoint is performed using multiple imputation method
  - Pattern of missing data was arbitrary so MCMC method used for imputation.

Method of Imputation

- MCMC: Performed by drawing simulations from Bayesian predication distribution for normal data

- Imputations done: Separately for each treatment group including baseline body weight (<90 kg, ≥90 kg) as a covariates

- The numbers of imputations were set to 500, seed number 4571234 for random function was used for this study. To generate the multiple imputed data sets, the SAS procedure MI was.

PSSI Score data – with missing observations

Unique Subject Identifier	Treatment	TRT(Numeric)	Weight category	weight_num	Baseline	LABEL OF FORMER VARIABLE	WEEK1	WEEK2	WEEK3	WEEK4	WEEK8	WEEK12
1012-1012005	Test drug	1	>= 90 Kg	0	21	Change from baseline	-6	-11	-15	-12	-20	-19
1020-1020002	Test drug	1	< 90 Kg	1	27	Change from baseline	.	-15	.	.	.	.
1021-1021003	Test drug	1	>= 90 Kg	0	60	Change from baseline	-5	-32	-40	-56	-58	-60
1021-1021011	Test drug	1	>= 90 Kg	0	24	Change from baseline	3	-15	-18	-24	-24	-24
1002-1002001	Placebo	2	< 90 Kg	1	45	Change from baseline	-13	-13	-33	-33	-33	-40
1002-1002006	Placebo	2	< 90 Kg	1	24	Change from baseline	-3	.	-6	.	.	.
1004-1004006	Placebo	2	>= 90 Kg	0	27	Change from baseline	-3	-3	-6	-9	-6	-3
1006-1006004	Placebo	2	< 90 Kg	1	45	Change from baseline	0	9	.	9	9	-3
1006-1006006	Placebo	2	< 90 Kg	1	54	Change from baseline	-19	-19	-14	-14	-14	-14

SAS Code

- Code
- options validvarname = upcase;
-
- proc mi data=data out = impdata1 seed = 4571234 nimpute = 500;
- var <weight_cat_num>< baseline> <week1 …… week12>;
- where trt in(1);
- run;
-
- proc mi data=data out = impdata1 seed = 4571234 nimpute = 500;
- var <weight_cat_num>< baseline> <week1 …… week12>;
- where trt in(2);
- run;

Output: sample data is showing the 500 imputation for one subject

Imputation Number	Unique Subject Identifier	Treatment	TRT	Weight category	weight_num	Baseline	WEEK1	WEEK2	WEEK3	WEEK4	WEEK8	WEEK12
1	1020-1020002	Test drug	1	< 90 Kg	1	27	-20.55	-15	-11.31	-7.707	-20.5	-18.35
2	1020-1020002	Test drug	1	< 90 Kg	1	27	-15.12	-15	-28.47	-22.86	-34.77	-33.36
3	1020-1020002	Test drug	1	< 90 Kg	1	27	-1.893	-15	-25.83	-27.32	-37.7	-44.64
4	1020-1020002	Test drug	1	< 90 Kg	1	27	-8.797	-15	-15.22	-21.86	-19.46	-25.57
5	1020-1020002	Test drug	1	< 90 Kg	1	27	-0.239	-15	-20.92	-16.63	-11.73	-13.83
..	....	.....	....	...	....	...	....	....	...	...	...	...
495	1020-1020002	Test drug	1	< 90 Kg	1	27	3.9565	-15	-19.82	-6.541	-8.272	-5.897
496	1020-1020002	Test drug	1	< 90 Kg	1	27	-7.61	-15	-22.39	-31.98	-28.52	-24.85
497	1020-1020002	Test drug	1	< 90 Kg	1	27	-19.37	-15	-10.24	-22.64	-19.22	-14.51
498	1020-1020002	Test drug	1	< 90 Kg	1	27	-8.249	-15	-18.3	-23.3	-16.36	-12.94
499	1020-1020002	Test drug	1	< 90 Kg	1	27	-4.578	-15	-19.41	-22.65	-38.66	-31.97
500	1020-1020002	Test drug	1	< 90 Kg	1	27	-7.721	-15	-20.5	-11.82	-15.69	-20.92

Analysis performed on imputed data

- After getting the above imputed data set we need to derive responder and non responder for PSSI 90.
- The PSSI 90 response variable was analyzed by Cochran-Mantel-Haenszel method with adjustment of body weight (<90 kg, ≥90 kg) and to get the 95% CI for the difference between the two treatment groups in the proportion of subjects for PSSI 90 responders were calculated using the normal approximation to the binomial distribution. Both analysis performed on 500 imputed data sets

SAS Code

```
proc freq data=impdata_t11;
                table <weight_cat_num *trtp*aval>/ cmh riskdiff;
                by _imputation_ avisitn;
                ods output CMH = t2 ; /* for pvalue: taking general association*/
run;

data t21(keep = _IMPUTATION_ AVISITN DF Prob);
                set t2;
                where AltHypothesis = "General Association ";
run;
```

- /*Statistical analysis: Binomial propoprtion difference and 95% CI*/
-
- proc freq data=impdata_t11;
- table trtpn*aval/ riskdiff;
- by _imputation_ avisitn ;
- ods output RiskDiffCol1=t6;
- run;

Result:

Imputation Number	AVISITN	DF	Prob	ESTIMATE	STDERR	95% Lower Confidence Limit	95% Upper Confidence Limit
1	2	1	0.0785	0.0588	0.0329	-0.0058	0.1234
1	3	1	0.0113	0.1176	0.0451	0.0292	0.2061
1	4	1	0.001	0.2157	0.0625	0.0932	0.3382
1	8	1	<.0001	0.3725	0.0711	0.2333	0.5118
1	12	1	<.0001	0.5098	0.0725	0.3676	0.652
2	2	1	0.0785	0.0588	0.0329	-0.0058	0.1234
2	3	1	0.0063	0.1373	0.0482	0.0428	0.2317
2	4	1	0.001	0.2157	0.0625	0.0932	0.3382
2	8	1	<.0001	0.3922	0.0716	0.2518	0.5325
2	12	1	<.0001	0.5294	0.0723	0.3876	0.6712
..	...	..	...	....	...	...	...
500	2	1	0.0785	0.0588	0.0329	-0.0058	0.1234
500	3	1	0.0113	0.1176	0.0451	0.0292	0.2061
500	4	1	0.001	0.2157	0.0625	0.0932	0.3382
500	8	1	<.0001	0.3725	0.0711	0.2333	0.5118
500	12	1	<.0001	0.5098	0.0725	0.3676	0.652

Inferential Step

- The result from the 500 complete dataset are then combined to produce inferential result . This was done by using Proc mianalyze and below code was used :

```
ods output parameterestimates = <estimate>;

proc mianalyze data = <final dataset>;

by visit;

modeleffects estimate;

stderr stderr;

run
```

After you analyze the m complete data sets by using standard SAS procedures. MIANALYZE procedure used to generate valid statistical inferences about parameters by combining results from the m analyses

Final result

A	B	C	D	E	F	G	H	I	J	K	L
AVISITN	Parameter	Estimate	Std Error	LCLMean	UCLMean	DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
1	estimate	0.019608	0.019415	.	.	.	0.019608	0.019608	0	.	.
2	estimate	0.058824	0.032948	.	.	.	0.058824	0.058824	0	.	.
3	estimate	0.121255	0.046326	0.0305	0.2121	684201	0.117647	0.137255	0	2.62	0.0089
4	estimate	0.221176	0.063641	0.0964	0.3459	1.20E+06	0.196078	0.235294	0	3.48	0.0005
8	estimate	0.379686	0.072018	0.2385	0.5208	1.38E+06	0.352941	0.392157	0	5.27	<.0001
12	estimate	0.514863	0.073582	0.3706	0.6591	915694	0.490196	0.529412	0	7	<.0001

Challenges:

- Issue: In the above data set for week 2 and week 3 between variability among the data set was zero (Due to all 500 estimates were constant) . So it was handled separately by taking average of p –values (prob)of treatment difference, Estimate and 95% CI and SE.
- Number of imputation required ?
- Check that data is following the characteristics of multivariate normal distribution

Recommendations

- No universally accepted approach
- Better to try to AVOID missing data up front
- Define in protocol how much is expected & how it will be dealt with. Be conservative.
- Do sensitivity analyses (check robustness of results)
- Discuss amount, implications etc. in study report.

References

- *Allison, P. D. (2001) Missing Data Thousand Oaks, CA: Sage Publications*
- *Cohen, J. & Cohen, P. (1983) Applied multiple regression/correlation analysis for the behavioral sciences (2nd ed.).Hillsdale, NJ: Erlbaum.*
- Weichung Joseph Shih, *Problems in dealing with missing data and informative censoring in clinical trials*

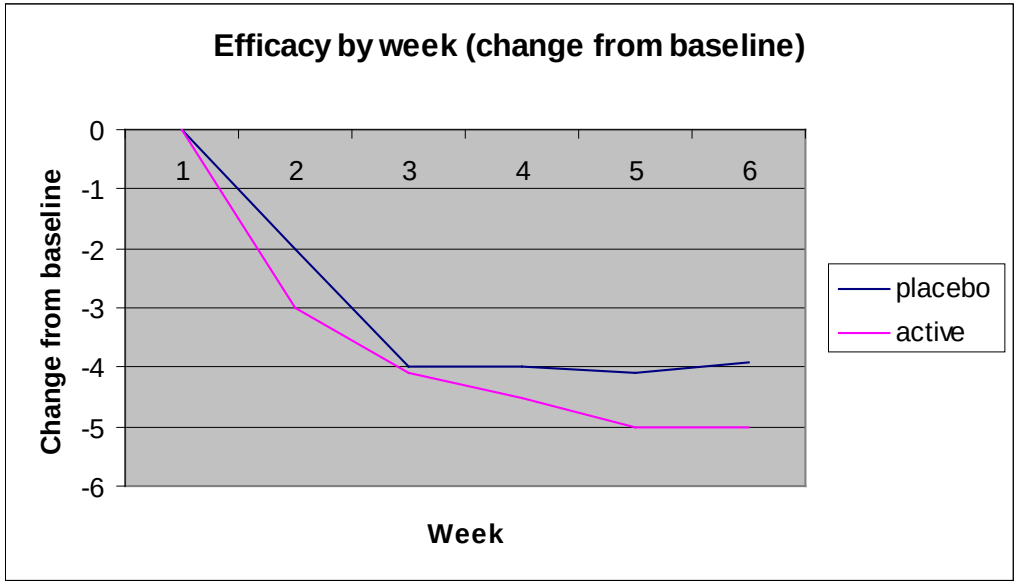
Thanks!

Cognizant

Appendix

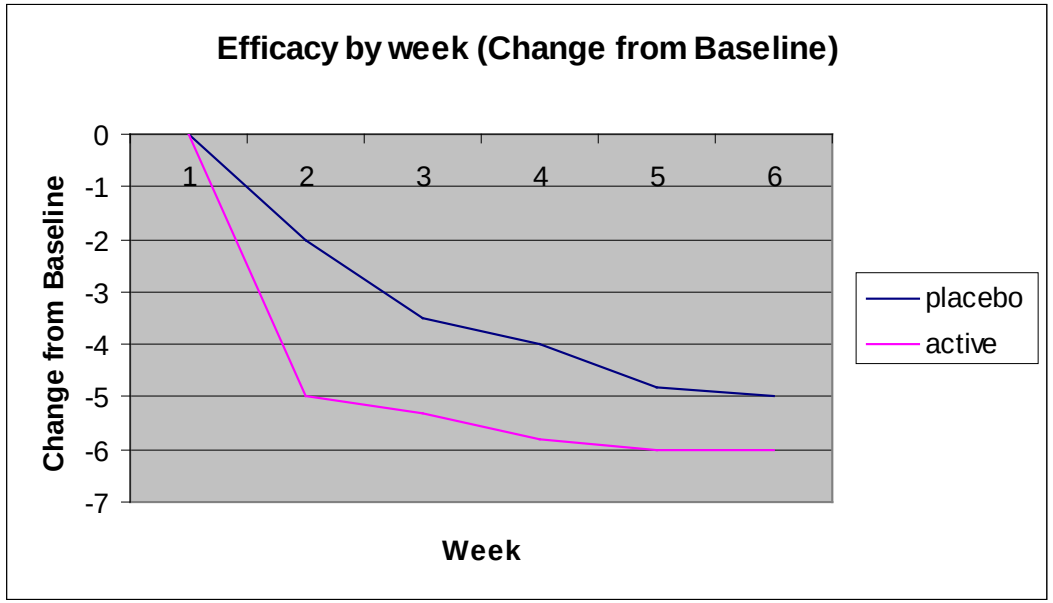
Example:

Scenario 1 : Placebo response levels off early



- LOCF is conservative
- Complete case analysis in not conservative

Scenario 1 : Scenario 2: Treatment effect levels off and placebo effect improves throughout



- LOCF is not conservative
- Complete case analysis is conservative

Result for Non Imputation

Statistical analysis (Cochran-Mantel-Haenszel test) of PSSI 90 response at Week 12 (non-responder imputation)
Full Analysis Set

Treatment comparison	Drug A 300 mg N=51 n/N' (%)	Placebo N=51 n/N' (%)	Difference between proportions (95% confidence interval)	P-value
Drug A 300 mg vs. Placebo	25/51 (52.9)	1/51 (2.0)	0.47 (0.35, 0.68)	<0.001

Patients with missing assessment were considered as non-responder.
n = number of patients with response.
N' = number of patients with evaluable data per treatment group.
P-value is calculated using Cochran-Mantel-Haenszel test adjusted for body weight (<90 kg, >=90 kg).
95% confidence interval for difference between proportions is calculated using the normal approximation to the binomial distribution.