




*«Проблема неполных данных:
способы, условия и методы
замещения данных»*

Вербицкая Елена Владимировна

доцент, кафедры клинической фармакологии и
доказательной медицины и
Зав. отделом фармакоэпидемиологии и биомедицинской
статистики

ПСПбГМУ им.акад. И.П.Павлова



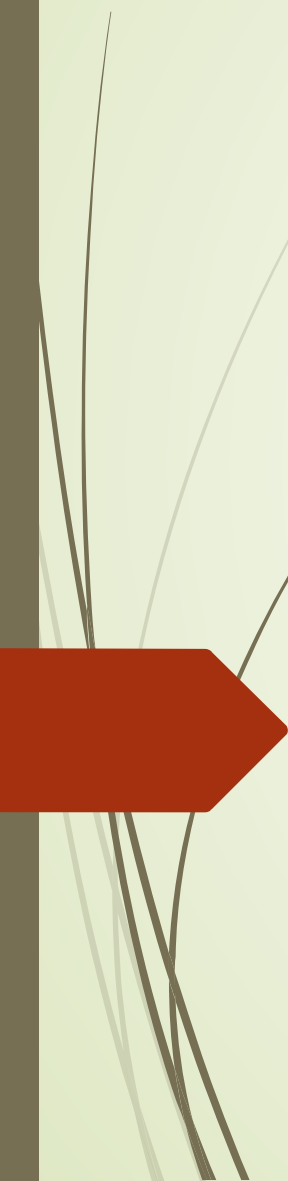
Проблемы с анализом данных с повторяющимися измерениями

- Сложные дизайны
- Много факторный анализ сложен для реализации
- **ПРОПУЩЕННЫЕ ДАННЫЕ**

Пример данных в наблюдательном исследовании

Группа 1 сайт1	Визит 1	170
	Визит 2	140
	Визит 3	46
	Визит 4	31
Группа 2 сайт1	Визит 1	206
	Визит 2	187
	Визит 3	17
	Визит 4	14

Группа 1 Сайт 2	Визит 1	379
	Визит 2	351
	Визит 3	0
	Визит 4	2
Группа 2 Сайт 2	Визит 1	45
	Визит 2	35
	Визит 3	0
	Визит 4	0



**Работа с
пропущенными
данными
Missing data analysis**



Losses-to-follow-up


How many is too many?

“5-and-20 rule of thumb”

- 5% probably leads to little bias
- >20% poses serious threats to validity


Depends on outcome event rate and comparative loss rates in the groups

Loss to follow-up rate should not exceed outcome event rate and should not be differential




Пропущенные данные – причины возникновения

- Истинно пропущенные данные –
 - Нет данных которые должны были быть получены (пробирки бьются, ответы теряются,)
 - Пациенты не приходят на визит
- Псевдо пропущенные данные
 - «Не приложимо» - беременности у мужчин



Пропущенные данные бывают разного типа

- missing completely at random (MCAR),
- missing at random (MAR)
- missing not at random (MNAR)



Пропущенные данные

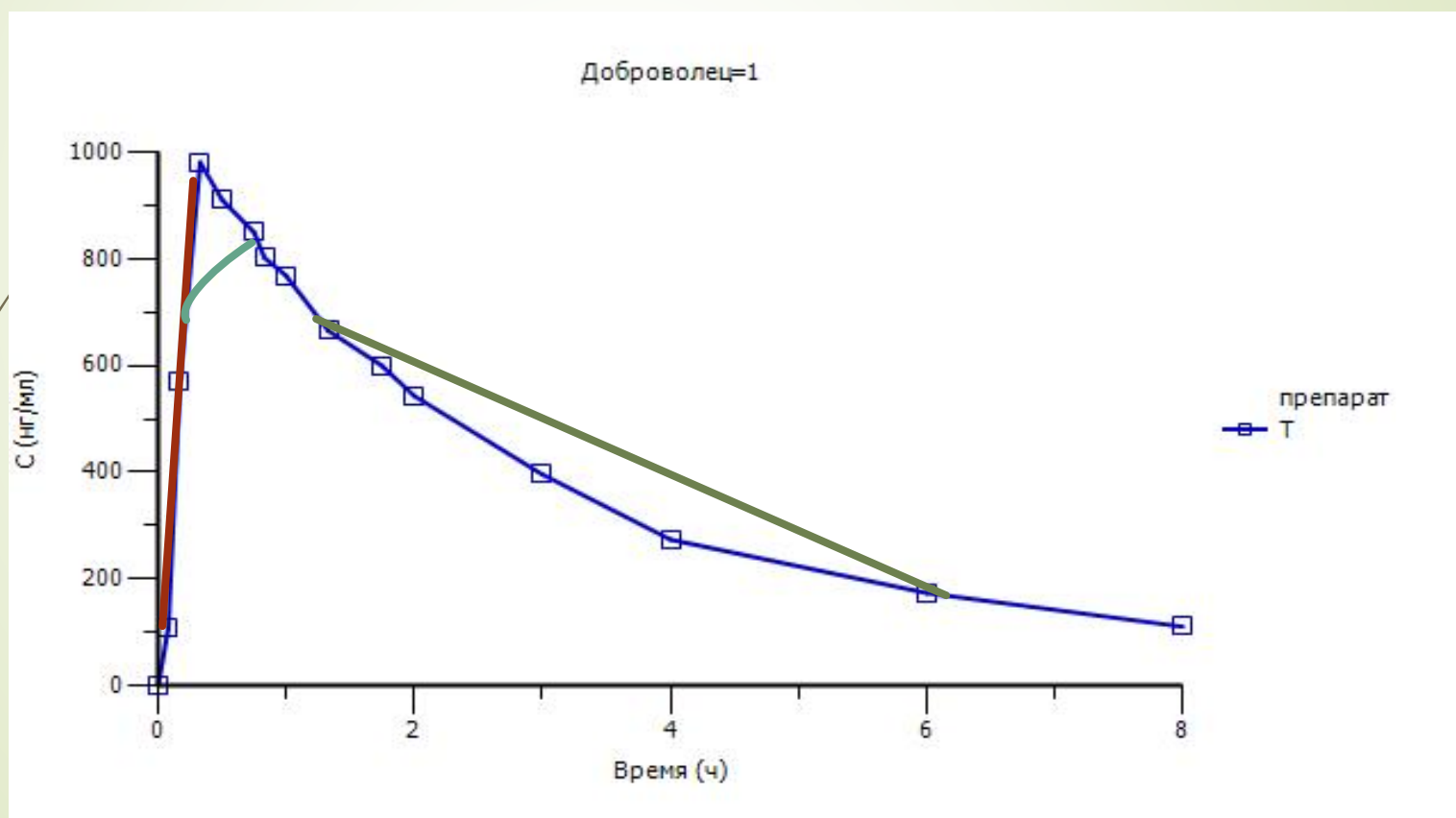
- ▶ MCAR представляет ситуацию, когда вероятность отсутствия наблюдения **не зависит от наблюдаемых или ненаблюдаемых данных.**
- ▶ MAR возникает, когда вероятность отсутствия наблюдения не зависит от ненаблюдаемых данных, **но она связана с некоторыми наблюдаемыми данными.**




Случайные пропущенные

- ▶ Оба случая предполагают случайность пропусков
 - ▶ MCAR – переехал, пробирка разбилась, потерялся анализ, на работе задержали, не смог доехать, подхватил простуду...
 - ▶ MAR – скорее всего случайно (однако, может зависеть от возраста или наличия какого-либо заболевания)
- ▶ Оба случая с определенной долей допущения могут игнорироваться ...

Случайные пропущенные





Метод работы с пропущенными данными

- Обработка только полных наблюдений
- Анализ доступных данных, без замещения пропущенных
- LOCF или другие «апостериорные методы замещения данных»
- Multiple imputation



Что в рекомендациях?

- ▶ четко не прописано, какие методы использовать
- ▶ «Unfortunately, no universally applicable methods of handling missing values can be recommended.» ICH E9
- ▶ В возможно использование в отдельных случаях (если прописано в протоколе) адекватных способов замещения данных



EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

2 July 2010
EMA/CPMP/EWP/1776/99 Rev. 1
Committee for Medicinal Products for Human Use (CHMP)

Guideline on Missing Data in Confirmatory Clinical Trials

4. The Effect of Missing Values on Data Analysis and Interpretation

4.1. Power and Variability

The sample size and the variability of the outcomes affect the power of a clinical trial. The power of a trial will increase if the sample size is increased or if the variability of the outcomes is reduced.

4.2. Bias

Bias is the most important concern resulting from missing data. If patients are excluded from the analysis, this may affect:


- The comparability of the treatment groups. A consequence of this may be a bias in the estimation of the treatment effect.
- The representativeness of the study sample in relation to the target population (external validity).



Обработка ТОЛЬКО ПОЛНЫХ наблюдений


- ▶ Пациент включается, если выполнил мин допустимое количество визитов (описанное в протоколе)
- ▶ Если доступны результаты по основному показателю (первичной точке)
- ▶ Отсутствие существенных отклонений от протокола, особенно на этапе включения

- ▶ Приводит к потере эффективности из-за недостатка данных
- ▶ Оценки слишком смещенные,
 - ▶ если данные не MCAR
 - ▶ Если пропущен существенный объем данных



Анализ доступных данных, без замещения пропущенных

- ▶ При небольшом числе пропусков
- ▶ Если данные MAR возможно получение не смещенных оценок
- ▶ «ignorable missingness» - только если адекватный анализ используется




Замена пропущенных данных

- ▶ Часто необходима для
 - ▶ анализа логитюдных данных (repeated measures ANOVA)
 - ▶ Построения регрессий
 - ▶

LOCF (LVCF)

- ▶ LOCF – замещение последним значимым результатом
 - ▶ Часто рекомендуется в клинических исследованиях, как валидный метод,
 - ▶ может как завышать значимость различий, так и снижать ее
 - ▶ Предполагается, что отражает применение препарата в бытовой практике



Другие «апостериорные методы замещения данных» (однократные замещения)

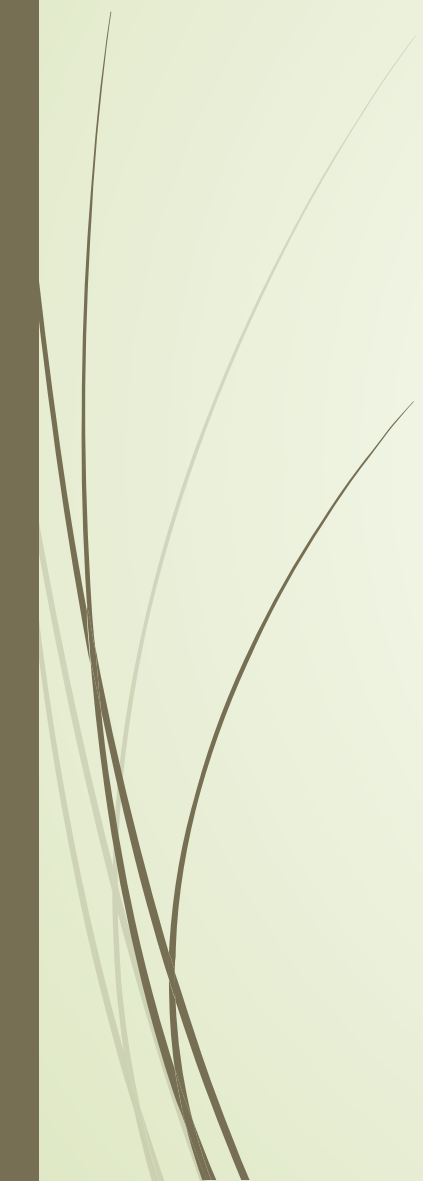
- ▶ Замена средней по группе
- ▶ Замена средней между двумя соседними значениям у одного человека,
- ▶ Замена случайно выбранным значением из данной группы

- ▶ **Замена значением на исходном уровне**

- ▶ **моделирование**



Multiple imputation

- ▶ На основе многократного непараметрического и предсказательного моделирования с учетом ранее полученных значений
 - ▶ Получают несколько «полных» баз данных
 - ▶ Анализируют каждый из вариантов
- 



Критика общепринятых МЕТОДОВ

- ▶ Missing data in randomised controlled trials — a practical guide //James R. Carpenter & Michael G. Kenward . November 21, 2007
- ▶ [Little RJ¹](#), [D'Agostino R](#), [Cohen ML](#), [Dickersin K](#), [Emerson SS](#), [Farrar JT](#), [Frangakis C](#), [Hogan JW](#), [Molenberghs G](#), [Murphy SA](#), [Neaton JD](#), [Rotnitzky A](#), [Scharfstein D](#), [Shih WJ](#), [Siegel JP](#), [Stern H](#).
The prevention and treatment of missing data in clinical trials. [N Engl J Med](#). 2012 Oct 4;367(14):1355-60. doi: 10.1056/NEJMSr1203730.
- ▶ Zhang, Yuqing et al. A systematic survey of the methods literature on the reporting quality and optimal methods of handling participants with missing outcome data for continuous outcomes in randomized controlled trials//Journal of Clinical Epidemiology , Volume 88 , 67 – 80
- ▶

RESEARCH ARTICLE

Open Access

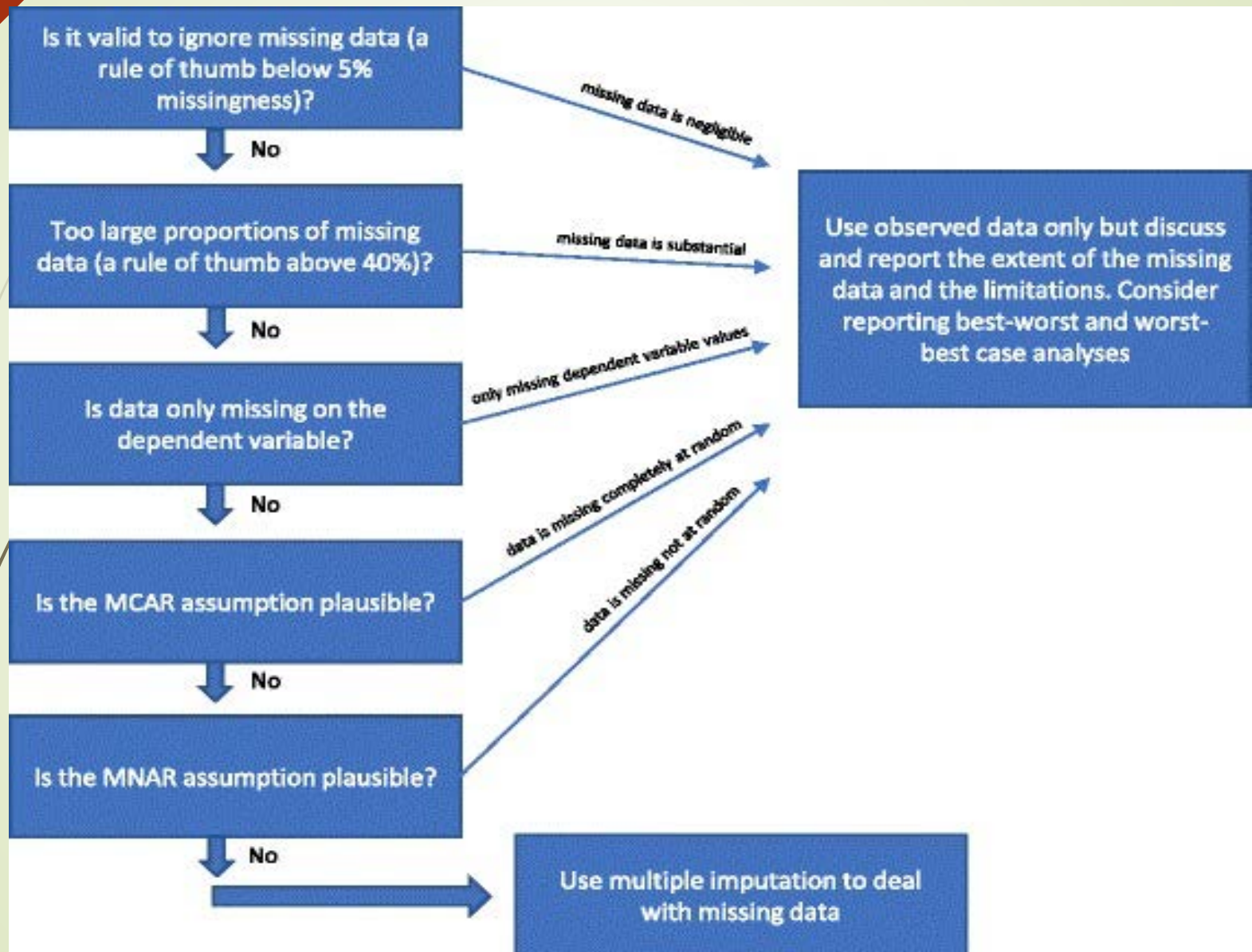



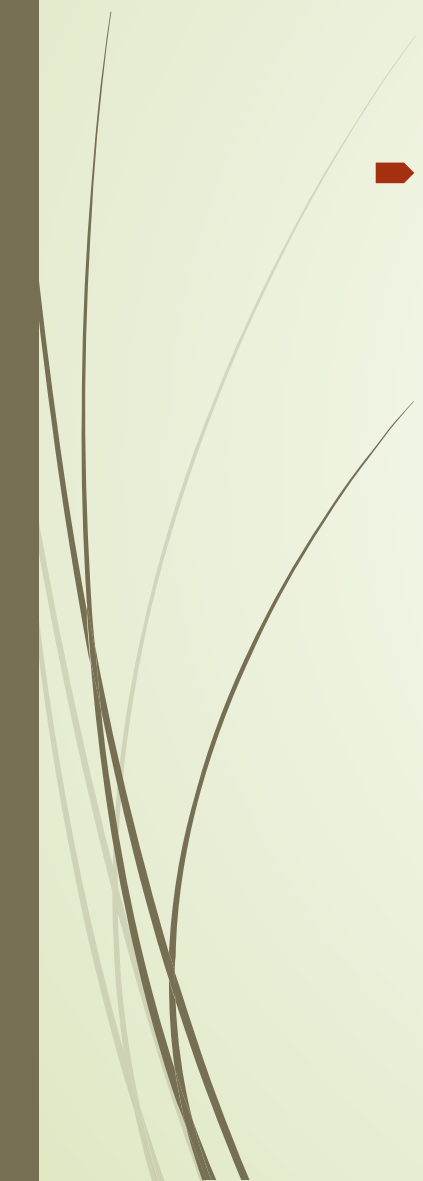
When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts

Janus Christian Jakobsen^{1,2*}, Christian Glud¹, Jørn Wetterslev¹ and Per Winkel¹

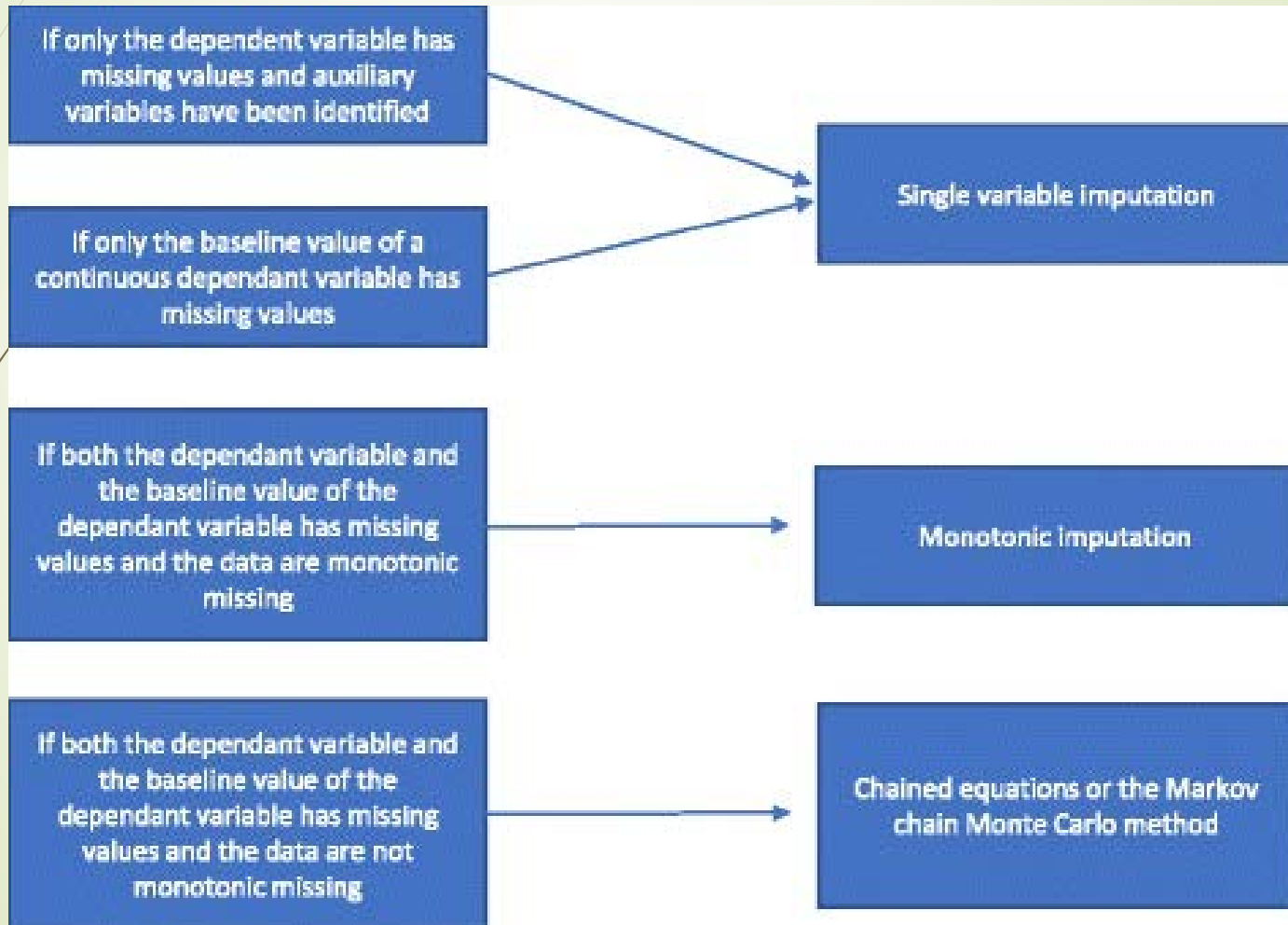
Jakobsen JC, Glud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. *BMC Medical Research Methodology*. 2017;17:162. doi:10.1186/s12874-017-0442-1.

КОГДА ИСПОЛЬЗОВАТЬ MI



- 
- 
- ▶ A worst case analysis: assigning the best possible outcome to missing values in the control group and the worst possible outcome to those of the experimental group. If this extreme analysis is still favorable then it can be confidently concluded that the results are robust to the handling of missing data.
 - ▶ - Анализ чувствительности рекомендуемый метод оценки влияния пропущенных данных
 - ▶ лучше всего работает на дихотомических данных
 - ▶ Сложно реализуется при количественных исходах


Какие замещения ИСПОЛЬЗОВАТЬ?



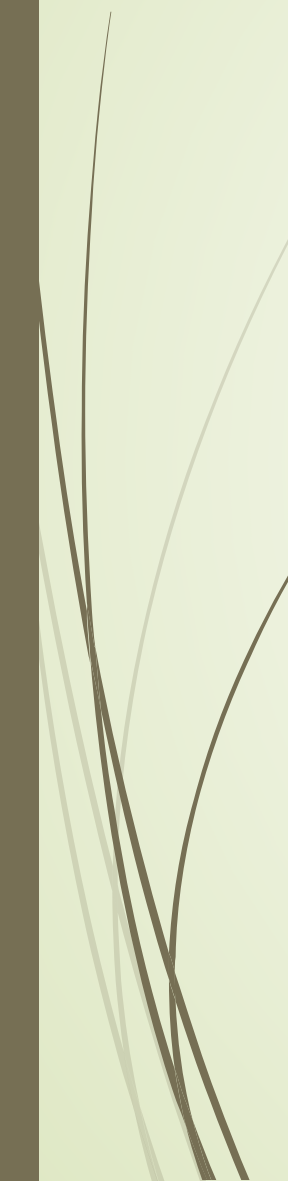


Не случайные пропуски

- ▶ **MNAR** means that after accounting for the available observed data, the reason for data being missing still depends on the unseen observations
- ▶ Причина пропуска зависит от пропущенных данных (неизвестных нам данных)
 - ▶ Пропустили по тому, что плохо себя чувствуют... Например, из-за развития побочных явлений, не нравится «качество оказания услуги»,



ДВОЙНОЕ СЛЕПОЕ
РАНДОМИЗИРОВАННОЕ ПЛАЦЕБО
КОНТРОЛИРУЕМОЕ ИССЛЕДОВАНИЕ
ПРИМЕНЕНИЯ ИМПЛАНТАТА
НАЛТРЕКСОНА (ПРОДЕТОКСОНА) ДЛЯ
ЛЕЧЕНИЯ ОПИЙНОЙ НАРКОМАНИИ



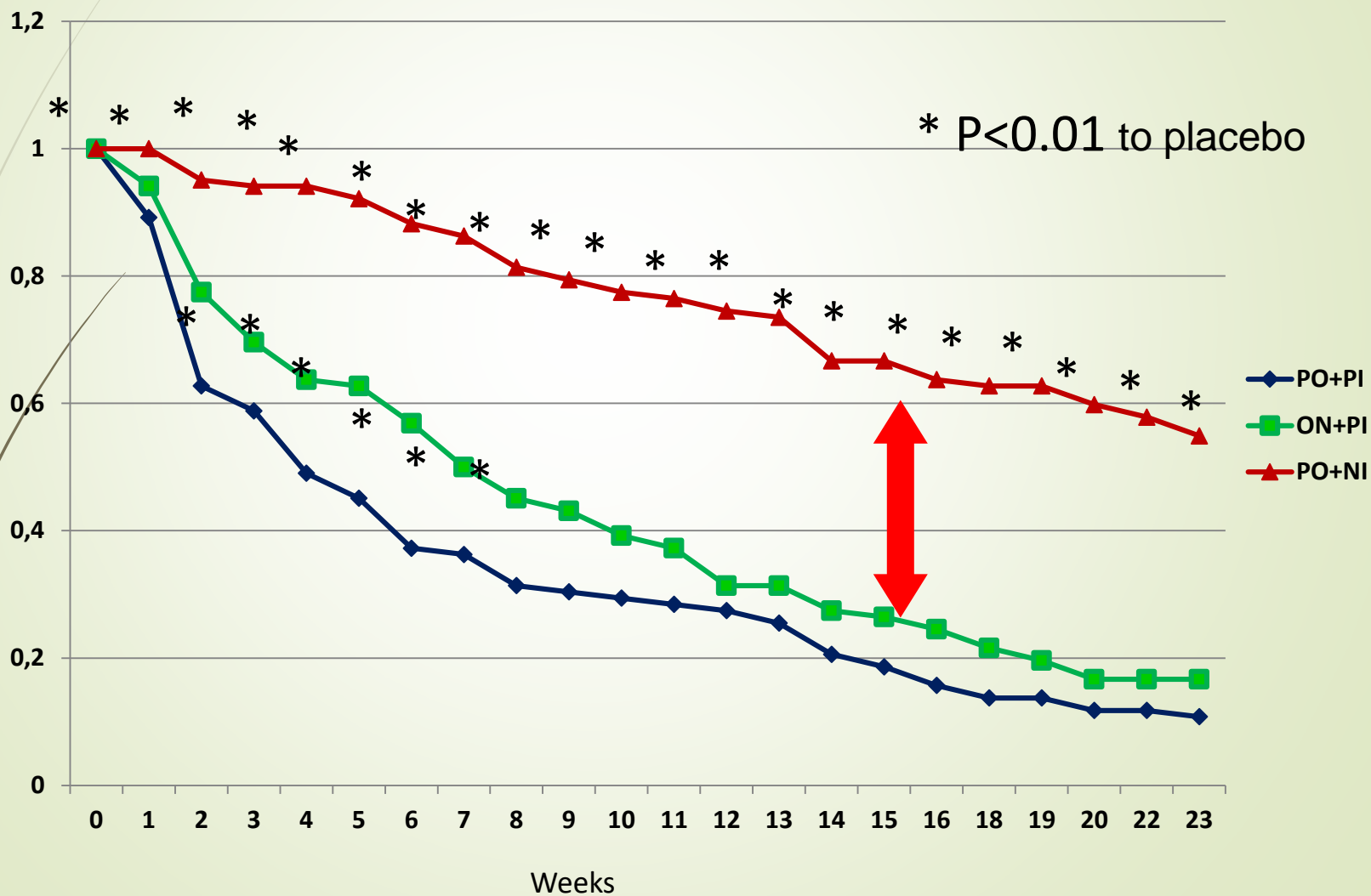
ГРУППЫ БОЛЬНЫХ

Показатели	Группы больных		
	OP+PI	ON+PI	OP+NI
Число больных	102	102	102
Пол (женщины)	27,5%	27,5%	27,5%
Возраст (M±SEM)	28,0±0,40	27,9±0,40	28,7±0,45
Давность наркомании (M±SEM)	7,8±0,38	7,9±0,41	8,3±0,39
Количество госпитализаций (M±SEM)	3,8±0,31	4,3±0,37	4,9±0,41

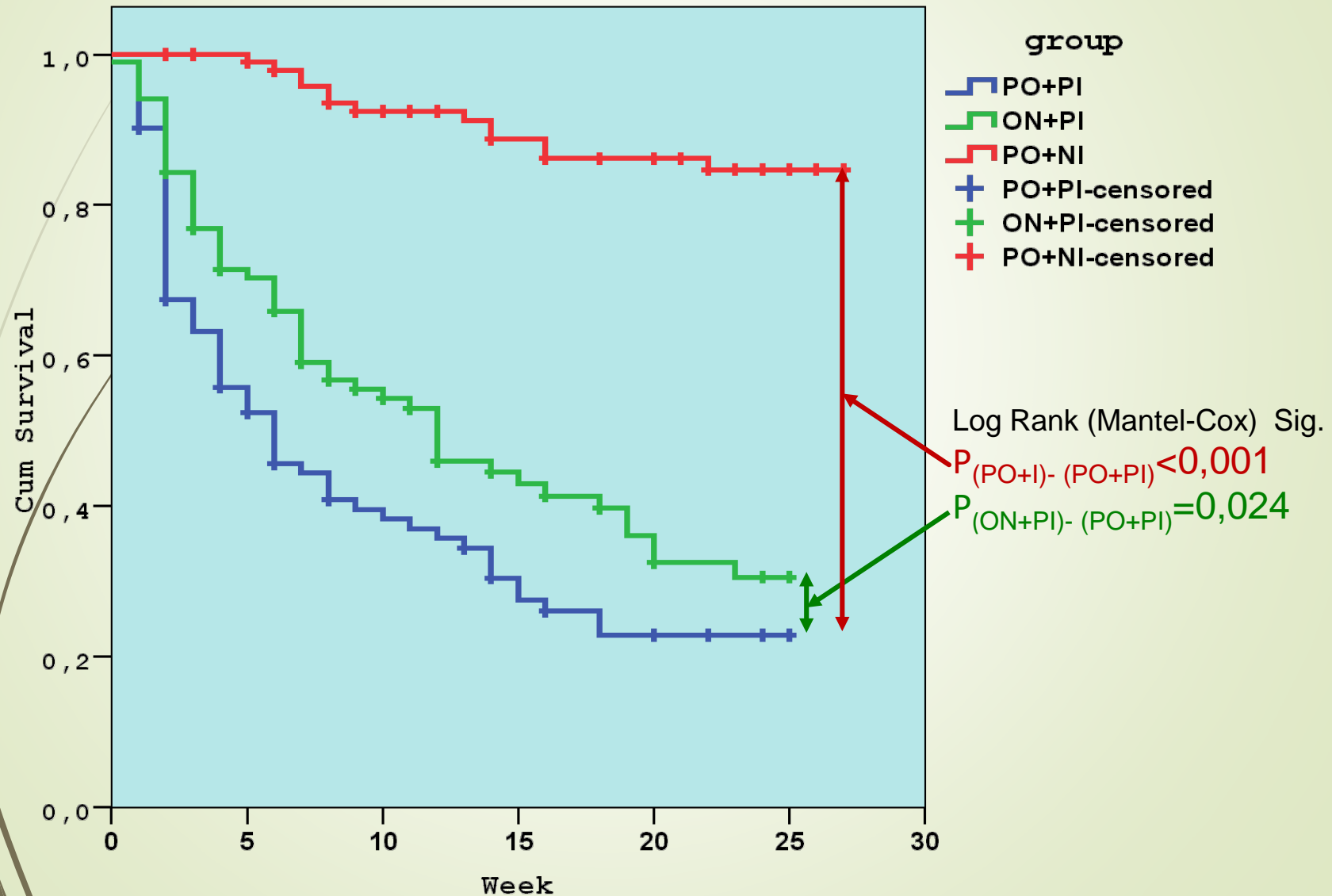
Показатели эффективности

- **Основные:**
 1. Выбывание из исследования и рецидив наркомании;
 2. Балл эффективности терапии (сумма героин-положительных и пропущенных анализов)
- **Вторичные показатели:** депрессия, тревога, ангедония, влечение к героину, риск ВИЧ-инфицирования, социальная адаптация .

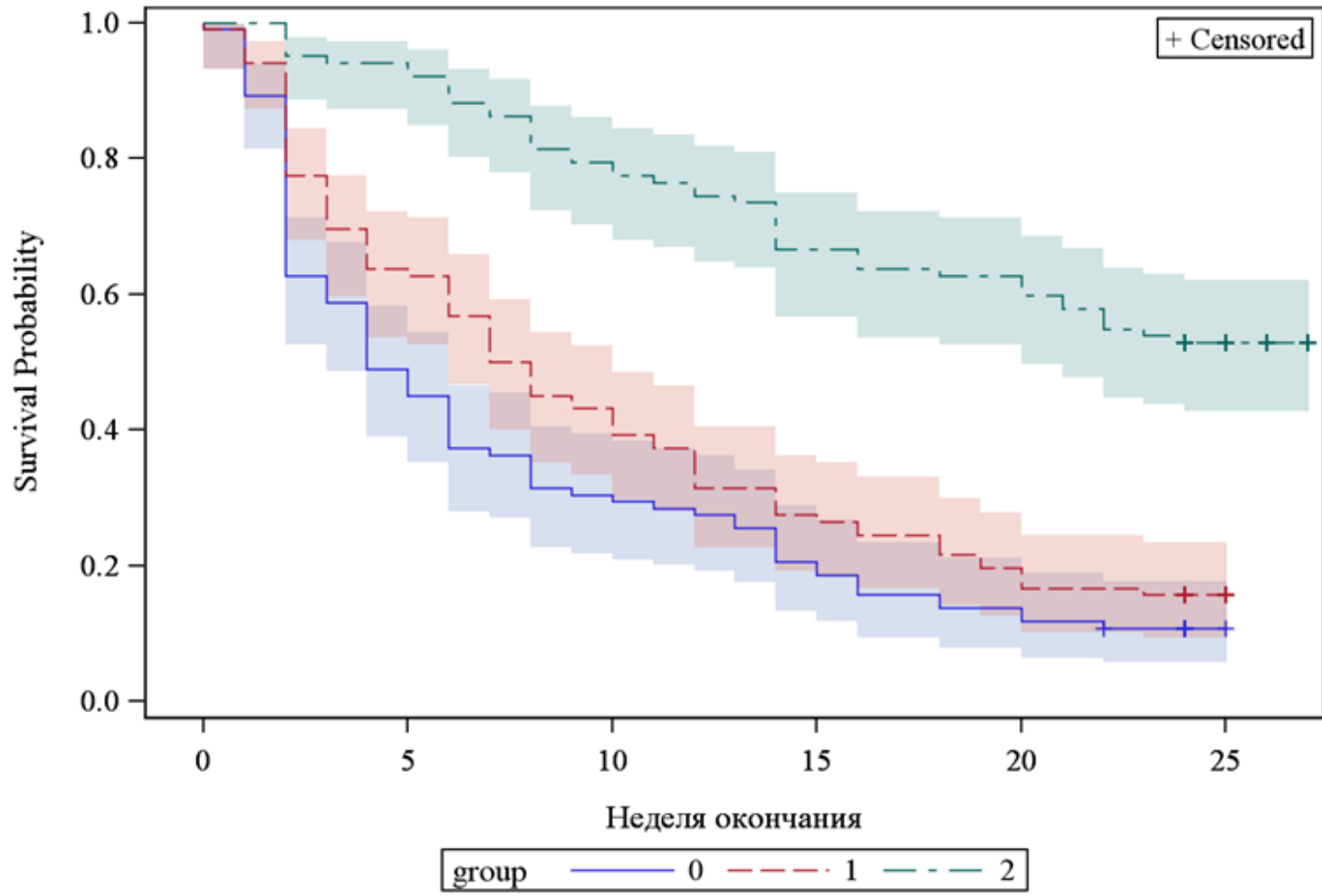
Удержание в программе (ремиссии) (%)



Kaplan-Meier Survival Functions

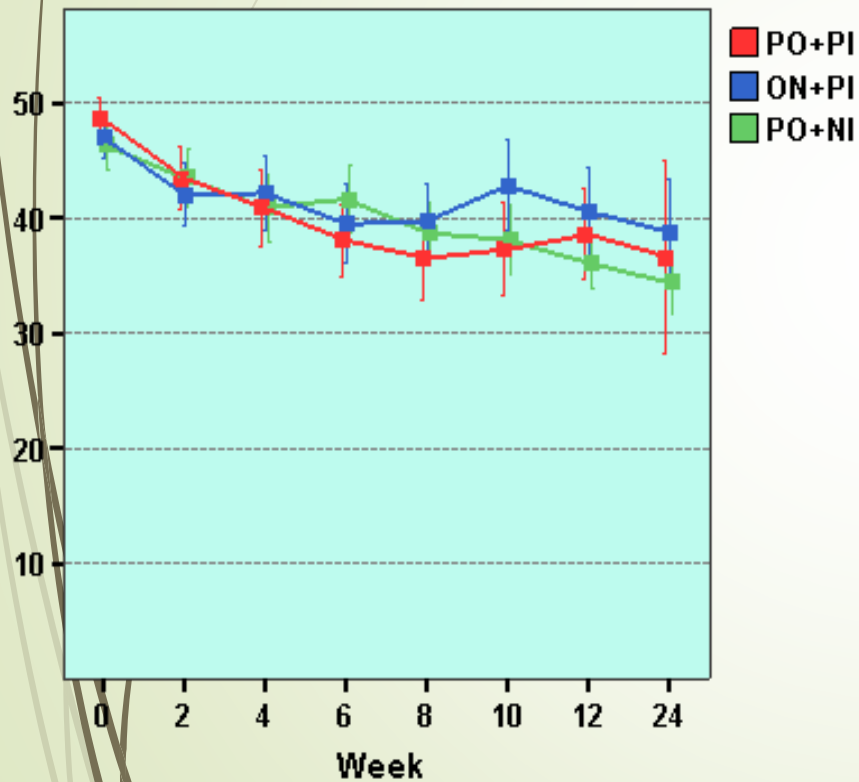


Product-Limit Survival Estimates With 95% Confidence Limits

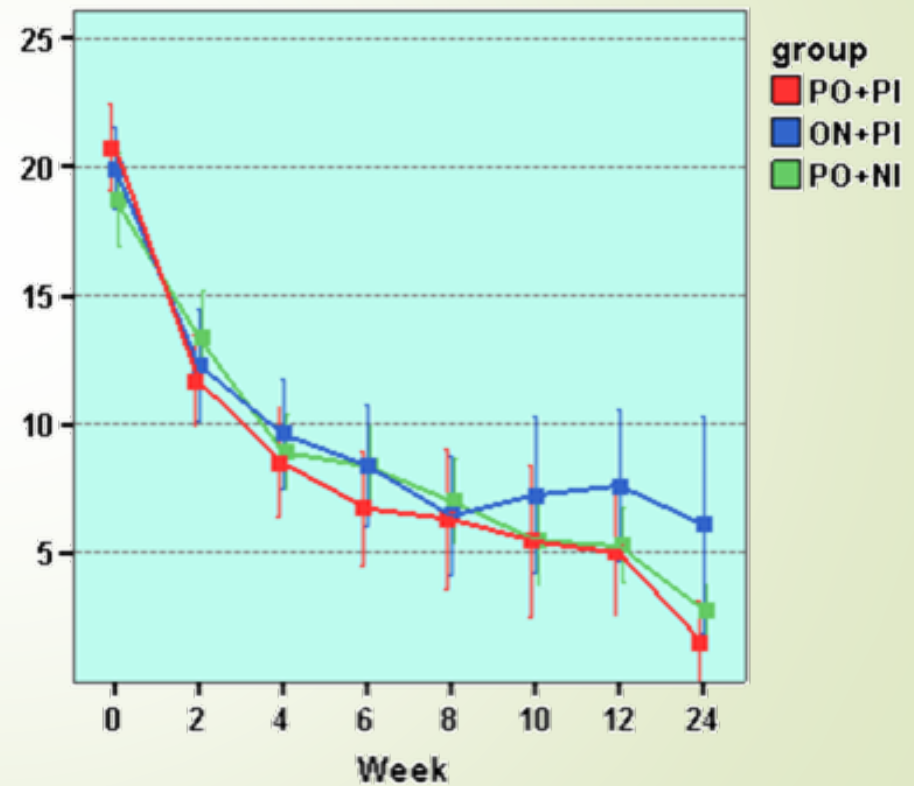


Тревога и Депрессия

Тревога (Шкала Спилбергера)



Депрессия (Шкала Бека)





Причины пропуска данных

- ▶ случайные или не случайные?
 - ▶ Рецидив – Каков уровень депрессии у того, кто сорвался?
 - ▶ Отсутствие контакта – Каков уровень депрессии у того, кто пропал из под наблюдения по другим причинам?
 - ▶ Связаны ли пропуски с исходным состоянием (длительностью зависимости, дозой героина, возрастом.....)
 - ▶

Являются ли пропущенные значения случайными?

1) Связь с известными переменными (MAR?)

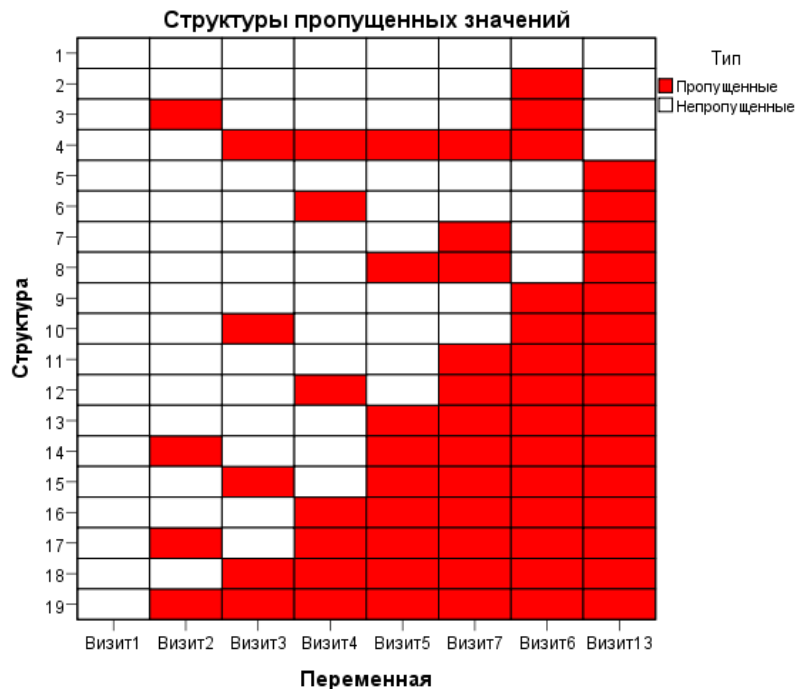
	Доза героина до	Длительность зависимости
Коэффициент корреляции ро Спирмена	0.299	0.046
Знч. (2-сторон)	<0.0001	0.528
N	192	192

2) Различия по количеству пропущенных между группами терапии

	Mean	SD	Min	Max	Q1	Median	Q3	U Манна-Уитни
PO+PI	4.5	2.39	0	7	3	5	7	P=0.0006
PO+NI	2.0	2.08	0	7	0	1	3	

Депрессия (Шкала Бека)

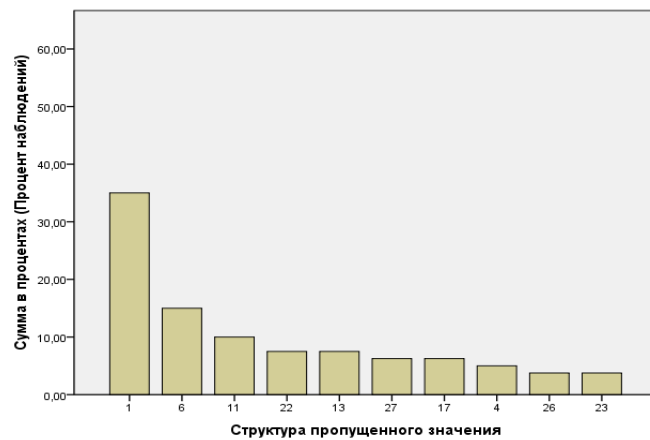
Плацебо



Продетоксон (имплант)

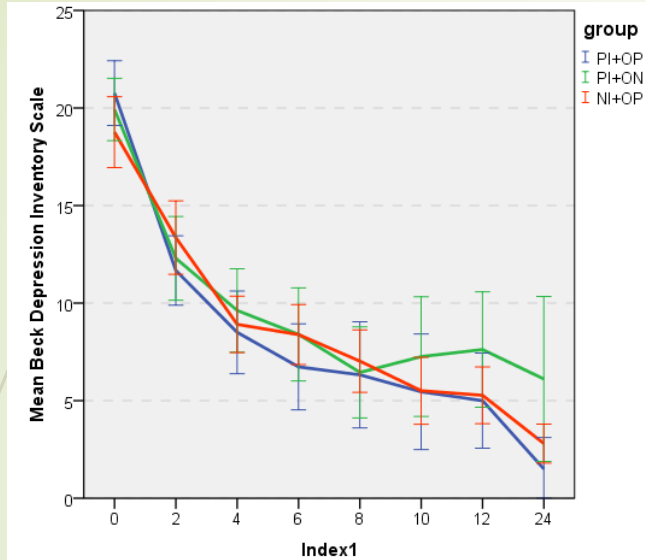


На диаграмме показаны 10 самых часто повторяющихся структур.

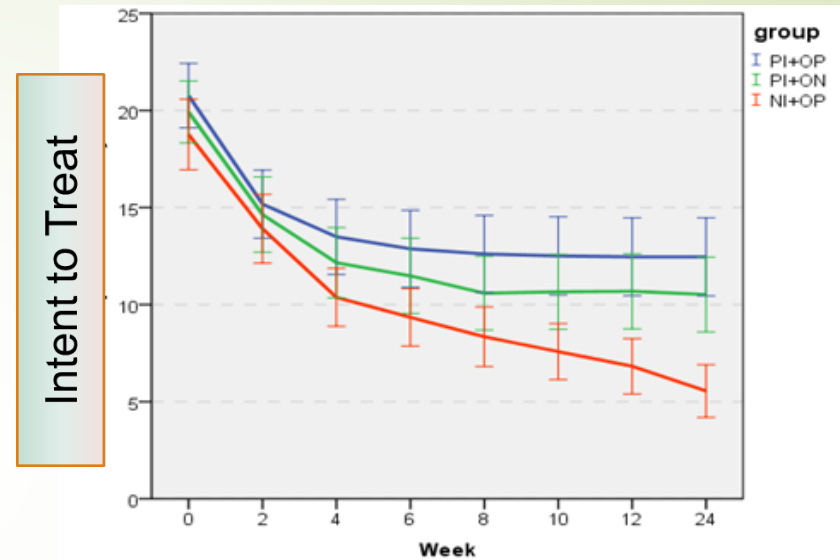


На диаграмме показаны 10 самых часто повторяющихся структур.

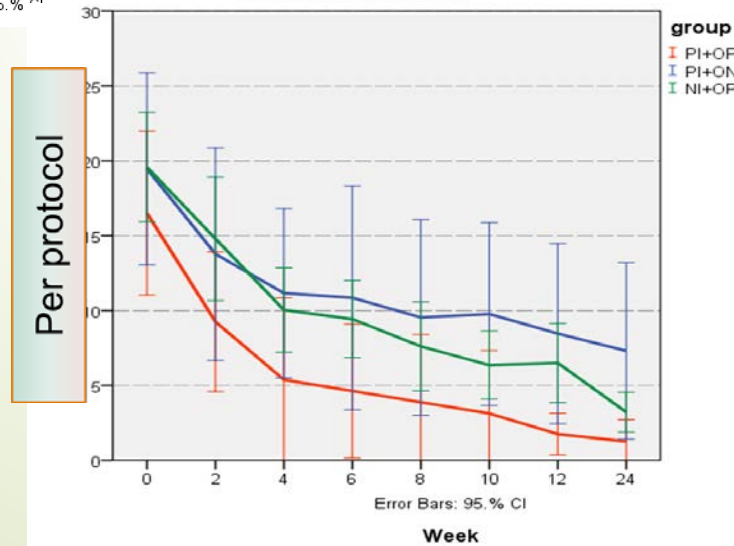
Депрессия (Шкала Бека)



Error Bars: 95.% CI

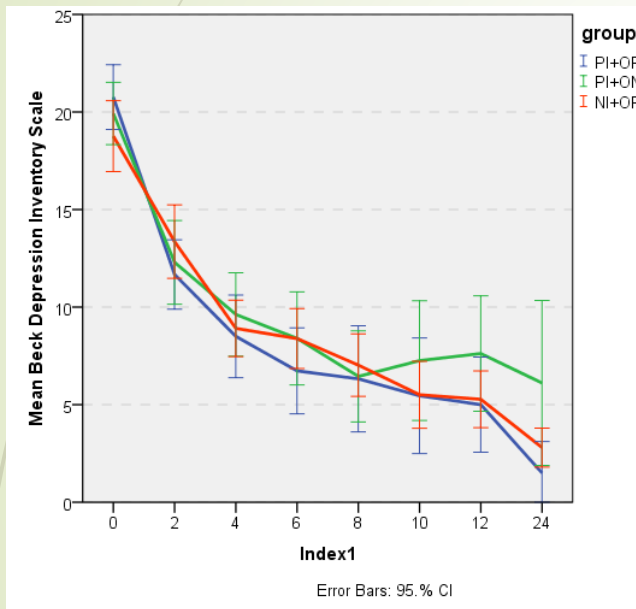


Error Bars: 95.% CI



Error Bars: 95.% CI

Анализ всех данных (сколько есть в каждой точке)



Проблемы:

- Недостаточно данных в последних точках
- Разное количество данных в конце исследования
- Разные дисперсии в группах в конце исследования

Алексеева Нина Петровна


- Разработан новый метод, решающий данные проблемы и корректно использующий все данные

(THE DEPENDENCE OF THE ERGODICITY ON THE TIME EFFECT IN THE REPEATED MEASURES ANOVA WITH MISSING DATA BASED ON THE UNBIASEDNESS RECOVERY Ufliand A., Alexeyeva N. В сборнике: [Springer Proceedings in Mathematics and Statistics](#) 7. 2014. С. 505-513.)

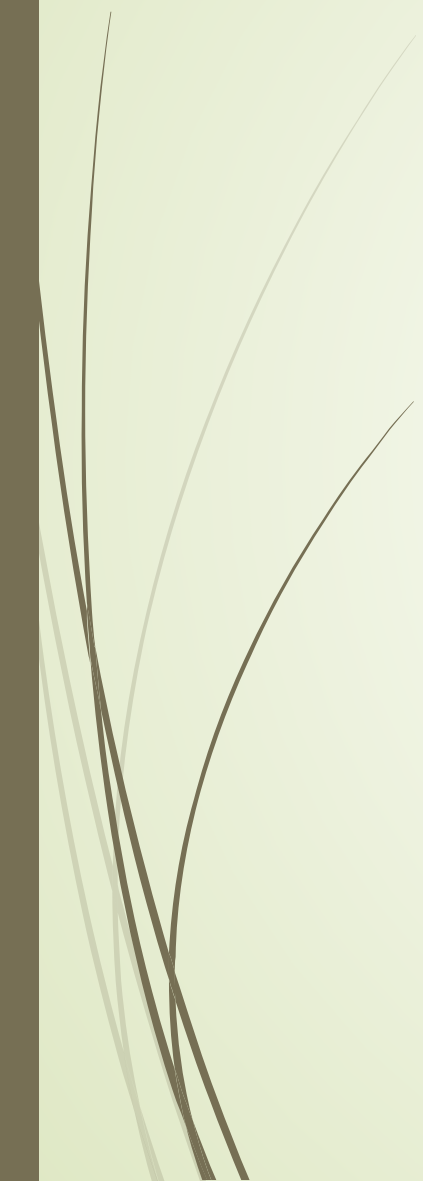


Анализ чувствительности?

- ▶ Каким значением замещать?
 - ▶ что считать "A worst case" ?
 - ▶ Сто считать "A best case" ?



Оценка значений состояния на
визите предшествующем
возобновлению приема опиатов



Endpoint (4 неделя)

Группы		Endpoint (4 неделя)				ANOVA
		Продолжают участие		Закончили программу		
		N	Mean±SE	N	Mean±SE	
Beck Depression						
2 weeks	PI+OP	64	19.5±1.07	38	23.0±1.39	Main group effect F2, 300=0.199; P=0.819 Main endpoint effect F1, 300=3.464; P=0.064 group*endpoint interaction F2, 300=0.066; P=0.07
	PI+ON	79	19.3±0.96	23	21.9±1.79	
	NI+OP	97	18.6±0.87	5	21.4±3.83	
Spielberger State Anxiety						
2 weeks	PI+OP	64	47.6±1.24	38	50.2±1.61	Main group effect F2, 300=0.264; P=0.768 Main endpoint effect F1, 300=5.906; P=0.016 group*endpoint interaction F2, 300=0.433; P=0.649
	PI+ON	79	46.2±1.12	23	49.8±2.07	
	NI+OP	97	46.0±1.01	5	53.2±4.44	
Spielberger Trait Anxiety						
2 weeks	PI+OP	64	47.3±1.16	38	50.5±1.51	Main group effect F2, 300=0.092; P=0.912 Main endpoint effect F1, 300=4.694; P=0.031 group*endpoint interaction F2, 300=0.037; P=0.963
	PI+ON	79	47.3±1.04	23	51.3±1.94	
	NI+OP	97	47.8±0.94	5	51.8±4.15	

Table 1. Eight Ideas for Limiting Missing Data in the Design of Clinical Trials.

Target a population that is not adequately served by current treatments and hence has an incentive to remain in the study.

Include a run-in period in which all patients are assigned to the active treatment, after which only those who tolerated and adhered to the therapy undergo randomization.

Allow a flexible treatment regimen that accommodates individual differences in efficacy and side effects in order to reduce the dropout rate because of a lack of efficacy or tolerability.

Consider add-on designs, in which a study treatment is added to an existing treatment, typically with a different mechanism of action known to be effective in previous studies.

Shorten the follow-up period for the primary outcome.

Allow the use of rescue medications that are designated as components of a treatment regimen in the study protocol.

For assessment of long-term efficacy (which is associated with an increased dropout rate), consider a randomized withdrawal design, in which only participants who have already received a study treatment without dropping out undergo randomization to continue to receive the treatment or switch to placebo.

Avoid outcome measures that are likely to lead to substantial missing data. In some cases, it may be appropriate to consider the time until the use of a rescue treatment as an outcome measure or the discontinuation of a study treatment as a form of treatment failure.

- Little R. J. et al. The Prevention and Treatment of Missing Data in Clinical Trials //The New England journal of medicine. – 2012. – T. 367. – №. 14. – C. 1355.



Спасибо за внимание!!!

Елена В.Вербичкая

К.б.н., доцент, отдел
фармакоэпидемиологии и
биомедицинской статистики