PHANDORIN: инструмент для контроля качества данных в нелинейном моделировании смешанных эффектов

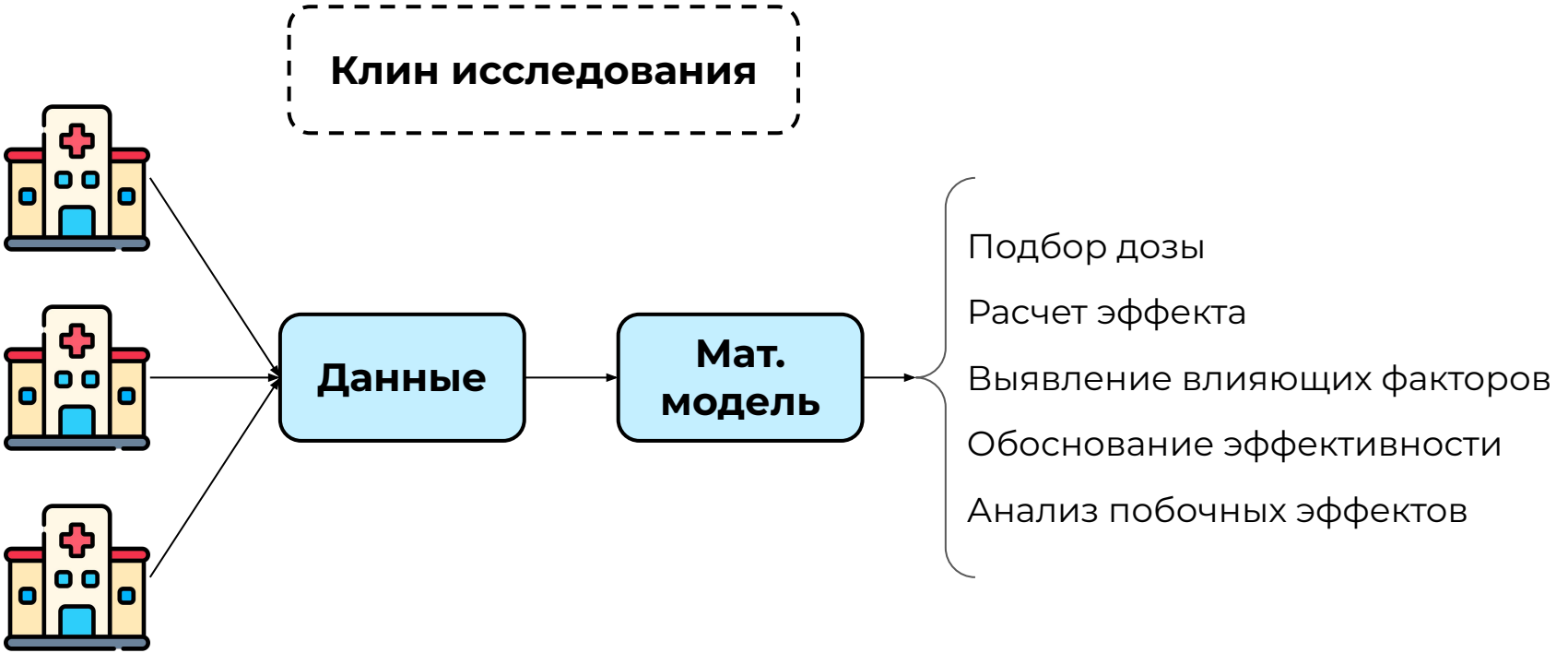
Никита Ваулин[1†], Игорь Васютин[2], Кирилл Жуденков[2]

[1] Сколтех

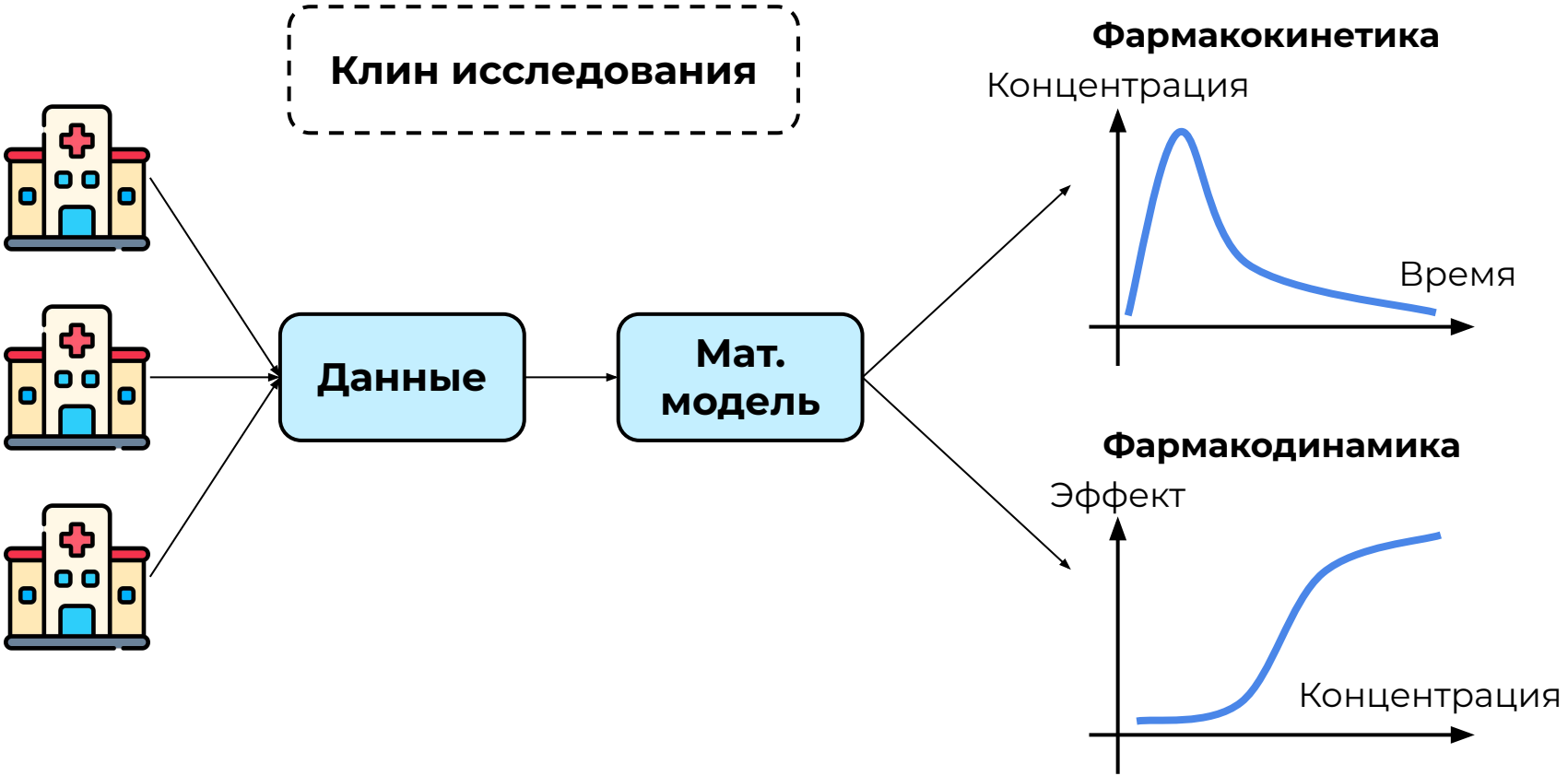
[2] M&S Decisions

[†]vaulin@ro.ru

Фармакометрика



Фармакометрика



Задача фармакокинетики (PK):

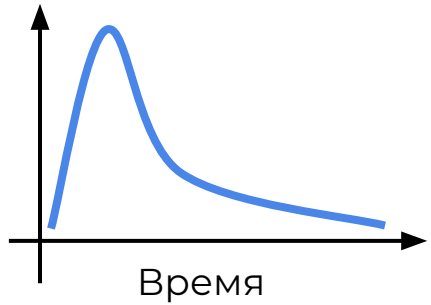
$$\frac{dA_d}{dt} = -k_a A_d$$

$$\frac{dA_c}{dt} = k_a A_d - k_e A_c$$

$$A_d|_{t=0} = Dose \qquad A_c|_{t=0} = 0$$

Концентрация

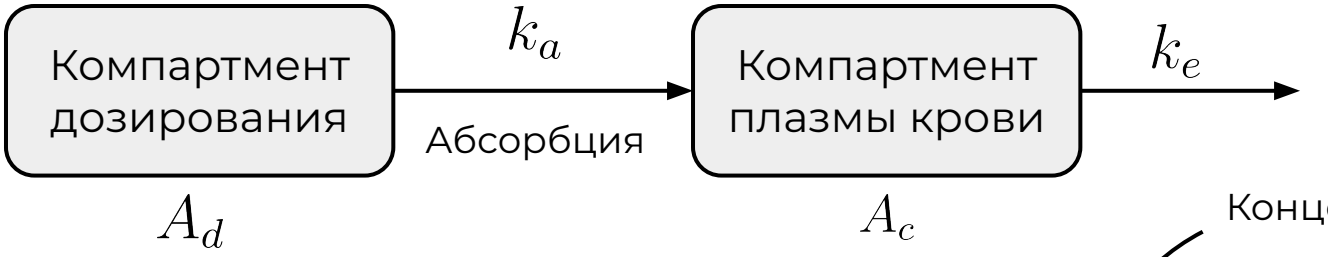
$$C_c = \frac{A_c}{V_c}$$



$$y_j = f(k_a, k_e, V) + \varepsilon_j$$

Решается обратная задача - оценка параметров (например, методом SAEM)

Задача фармакокинетики (PopPK):



$$\frac{dA_d}{dt} = -k_a A_d$$

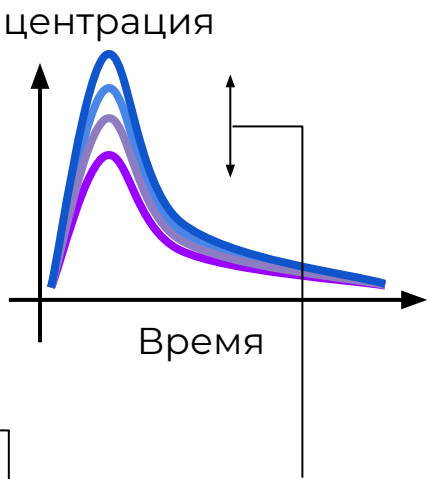
$$\frac{dA_c}{dt} = k_a A_d - k_e A_c$$

$$A_d|_{t=0} = Dose \qquad A_c|_{t=0} = 0$$

$$C_c = \frac{A_c}{V_c}$$

~~$y_j = f(k_a, k_e, V) + \varepsilon_j$~~

$$y_{ij} = f(k_{a_i}, k_{e_i}, V_i) + \varepsilon_{ij}$$
$$V_i = V + \eta_i$$

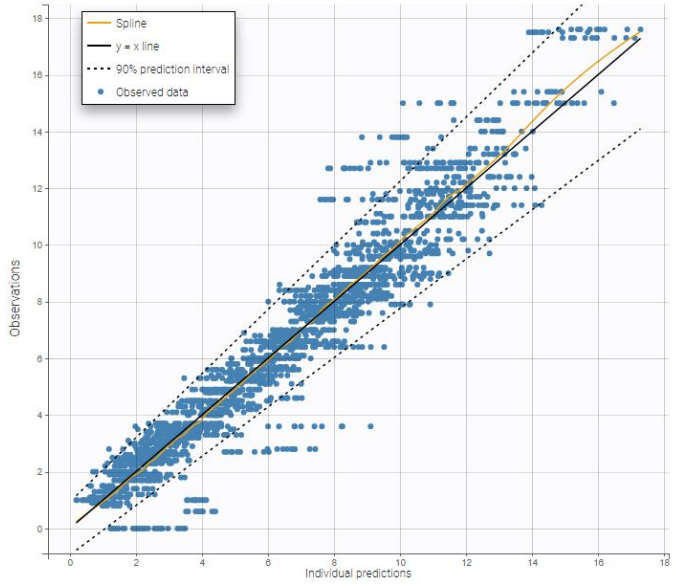


ПО для моделирования

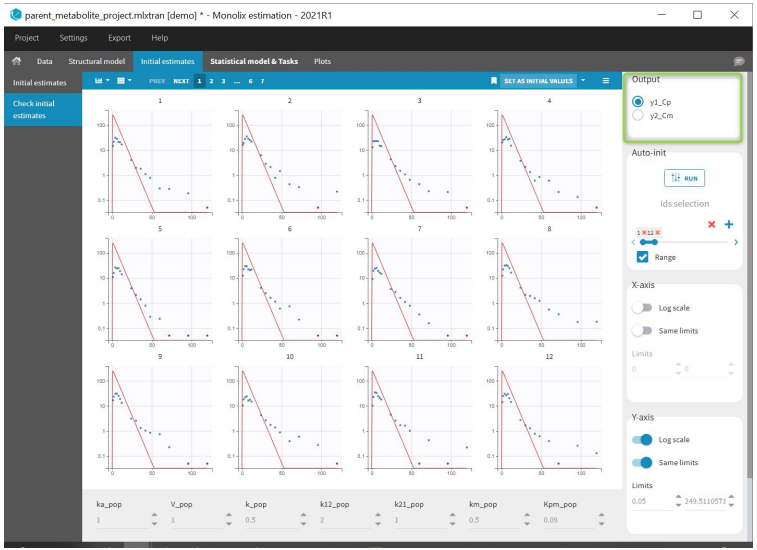
Пакеты для R

Отдельное ПО

nlmixr²



Monolix



Особенный формат данных

	CATEGORICAL COVARIATE	ID	EVENT ID	TIME	AMOUNT	IGNORE	OBSERVATION	CENSORING	LIMIT	IGNORED OBSERVATION	IGNORE	IGNORE	IGNORE	CONTINUOUS COVARIATE	CATEGORICAL COVARIATE	CATEGORICAL COVARIATE	IGNORE	IGNORE	CONTINUOUS COVARIATE
LINE NUMBER	STUDY	ID	EVID	TIME	AMT	DUR	DV	BLQ	LIMIT	MDV	DGROUP	BWT	BBSA	AGE	SEX	RACE	BILI	CRCL	HT
2	S1	101	1	0	2.2	0.5	.	.	.	0	1	98	2.19	55	1	1	0.4	67.1	185
3	S1	101	0	0.35	.	.	15.8	0	0	0	1	98	2.19	55	1	1	0.4	67.1	185
4	S1	101	0	0.49	.	.	16.8	0	0	0	1	98	2.19	55	1	1	0.4	67.1	185
5	S1	101	0	0.74	.	.	6.9	0	0	0	1	98	2.19	55	1	1	0.4	67.1	185
6	S1	101	0	1	.	.	4.4	0	0	0	1	98	2.19	55	1	1	0.4	67.1	185
7	S1	101	0	1.55	.	.	3.1	0	0	0	1	98	2.19	55	1	1	0.4	67.1	185
8	S1	101	0	2	.	.	2.7	0	0	0	1	98	2.19	55	1	1	0.4	67.1	185
9	S1	101	0	3	.	.	2.5	1	0	0	1	98	2.19	55	1	1	0.4	67.1	185
10	S1	101	0	4	.	.	2.5	1	0	0	1	98	2.19	55	1	1	0.4	67.1	185
11	S1	101	0	5.83	.	.	2.5	1	0	0	1	98	2.19	55	1	1	0.4	67.1	185
12	S1	101	0	8.08	.	.	2.5	1	0	0	1	98	2.19	55	1	1	0.4	67.1	185
13	S1	101	0	24.05	.	.	2.5	1	0	0	1	98	2.19	55	1	1	0.4	67.1	185

... 10-ки тысяч записей

Особенный формат данных

	CATEGORICAL COVARIATE	ID	EVENT ID	TIME	AMOUNT	IGNORE	OBSERVATION	CENSORING	LIMIT	IGNORED OBSERVATION	IGNORE	IGNORE	IGNORE	CONTINUOUS COVARIATE	CATEGORICAL COVARIATE	CATEGORICAL COVARIATE	IGNORE	IGNORE	CONTINUOUS COVARIATE
LINE NUMBER	STUDY	ID	EVID	TIME	AMT	DUR	DV	BLQ	LIMIT	MDV	DGROUP	BWT	BBSA	AGE	SEX	RACE	BILI	CRCL	HT
2	S1	101	1	0	2.2	0.5	.	.	.	0	1	98	2.19	55	1	1	0.4	67.1	185
3	S1	101	0	0.35	.	.	15.8	0	0	0	1	98	2.19	55	1	1	0.4	67.1	185
4	S1	101	0	0.49	.	.	16.8	0	0	0	1	98	2.19	55	1	1	0.4	67.1	185
5	S1	101	0	0.74	.	.	6.9	0	0	0	1	98	2.19	55	1	1	0.4	67.1	185
6	S1	101	0	1	.	.	4.4	0	0	0	1	98	2.19	55	1	1	0.4	67.1	185
7	S1	101	0	1.55	.	.	3.1	0	0	0	1	98	2.19	55	1	1	0.4	67.1	185
8	S1	101	0	2	.	.	2.7	0	0	0	1	98	2.19	55	1	1	0.4	67.1	185
9	S1	101	0	3	.	.	2.5	1	0	0	1	98	2.19	55	1	1	0.4	67.1	185
10	S1	101	0	4	.	.	2.5	1	0	0	1	98	2.19	55	1	1	0.4	67.1	185
11	S1	101	0	5.83	.	.	2.5	1	0	0	1	98	2.19	55	1	1	0.4	67.1	185
12	S1	101	0	8.08	.	.	2.5	1	0	0	1	98	2.19	55	1	1	0.4	67.1	185
13	S1	101	0	24.05	.	.	2.5	1	0	0	1	98	2.19	55	1	1	0.4	67.1	185

Особенный формат данных

ID	EVID	TIME	AMT	DUR	DV
101	1	0	2.2	0.5	.
101	0	0.35		.	15.8
101	0	0.49	.	.	16.8
101	0	0.74	.	.	6.9
101	0	1	.	.	4.4
101	0	1.55		.	3.1

ID	Идентификатор пациента
EVID	Тип записи
TIME	Время
AMOUNT	Доза
DV	Зависимая переменная (наблюдение)
Ковариаты: численные или категориальные, постоянные или изменяющиеся	
Дополнительные специальные колонки	

Особенный формат данных

ID	EVID	TIME	AMT	DUR	DV
101	1	0	2.2	0.5	.
101	0	0.35	.	.	15.8
101	0	0.49	.	.	16.8
101	0	0.74	.	.	6.9
101	0	1	.	.	4.4
101	0	1.55	.	.	3.1

Была дана доза препарата

ID	Идентификатор пациента
EVID	Тип записи
TIME	Время
AMOUNT	Доза
DV	Зависимая переменная (наблюдение)
Ковариаты: численные или категориальные, постоянные или изменяющиеся	
Дополнительные специальные колонки	

Особенный формат данных

ID	EVID	TIME	AMT	DUR	DV
101	1	0	2.2	0.5	.
101	0	0.35	.	.	15.8
101	0	0.49	.	.	16.8
101	0	0.74	.	.	6.9
101	0	1	.	.	4.4
101	0	1.55	.	.	3.1

Записываются наблюдение
(снятие зависимой переменной)

ID	Идентификатор пациента
EVID	Тип записи
TIME	Время
AMOUNT	Доза
DV	Зависимая переменная (наблюдение)
Ковариаты: численные или категориальные, постоянные или изменяющиеся	
Дополнительные специальные колонки	

Особенный формат данных

ID	EVID	TIME	AMT	DUR	DV
101	1	0	2.2	0.5	.
101	0	0.35			15.8
101	0	0.49			16.8
101	0	0.74			6.9
101	0	1			4.4
101	0	1.55			3.1

Некоторые другие колонки также зависят друг от друга

ID	Идентификатор пациента
EVID	Тип записи
TIME	Время
AMOUNT	Доза
DV	Зависимая переменная (наблюдение)
Ковариаты: численные или категориальные, постоянные или изменяющиеся	
Дополнительные специальные колонки	

Проблема контроля качества данных

В фармакометрике данные:

- Имеют очень **специфическую структуру**
- Собраны из **множества исследований**

} Повышен риск возникновения ошибок

В то же время процедура контроля качества:

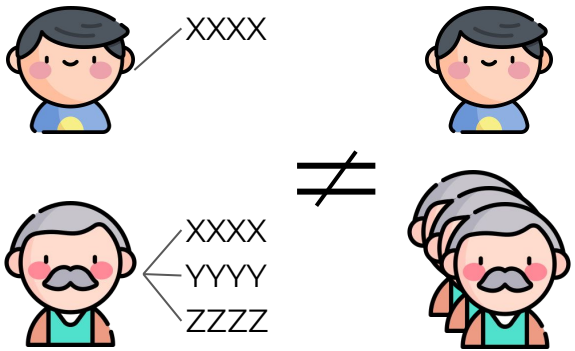
- Не унифицирована
- Монотонна и рутинна
- В имеющемся ПО оставлена на этап анализа модели

} Повышен риск пропуска ошибок

Типичные проблемы

Разное число записей на пациента

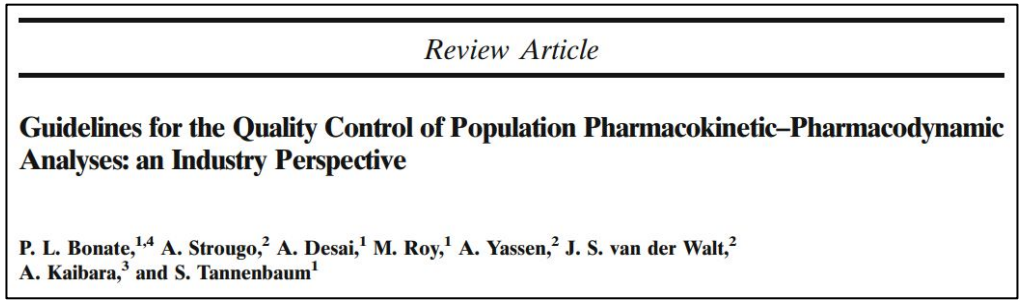
ID	Время	Доза
1	0	100
1	0	100
2	0	100



Внесение записей в разных размерностях

ID	Время	Доза
1	0	10
2	0	10
3	0	**1000**

Проблема контроля качества данных

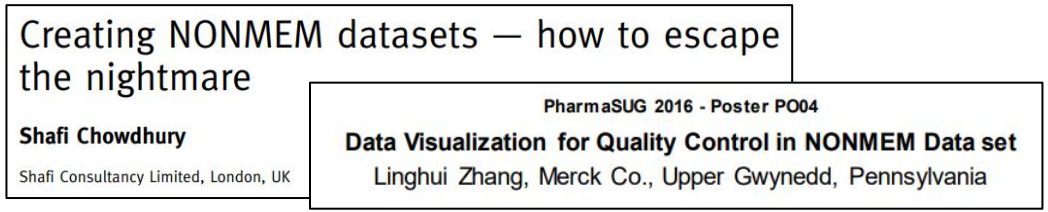


Review Article

Guidelines for the Quality Control of Population Pharmacokinetic–Pharmacodynamic Analyses: an Industry Perspective

P. L. Bonate,[1,4] A. Strougo,[2] A. Desai,[1] M. Roy,[1] A. Yassen,[2] J. S. van der Walt,[2] A. Kaibara,[3] and S. Tannenbaum[1]

- Разбиение процедуры QC на отдельные блоки (ковариаты, наблюдения технические ошибки, …)
- Подробный опросник для QC



Creating NONMEM datasets — how to escape the nightmare

Shafi Chowdhury

Shafi Consultancy Limited, London, UK

PharmaSUG 2016 - Poster PO04

Data Visualization for Quality Control in NONMEM Data set

Linghui Zhang, Merck Co., Upper Gwynedd, Pennsylvania

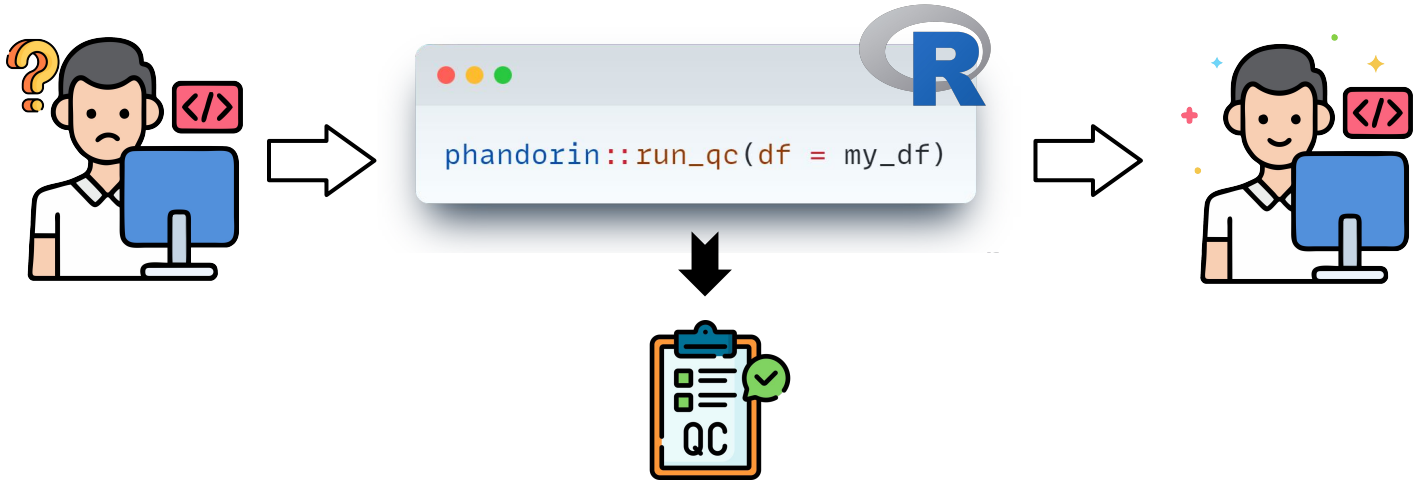
- Дополнительные вопросы для QC
- Способы отображения данных

Bonate, P. L., et al. "Guidelines for the quality control of population pharmacokinetic–pharmacodynamic analyses: an industry perspective." The AAPS journal 14 (2012): 749-758;
Chowdhury, Shafi. "Creating NONMEM datasets—how to escape the nightmare." Pharmaceutical programming 3.2 (2010): 80-83;
Zhang, L., et al. " Data Visualization for Quality Control in NONMEM Data set." PharmaSUG Poster PO04 (2016);

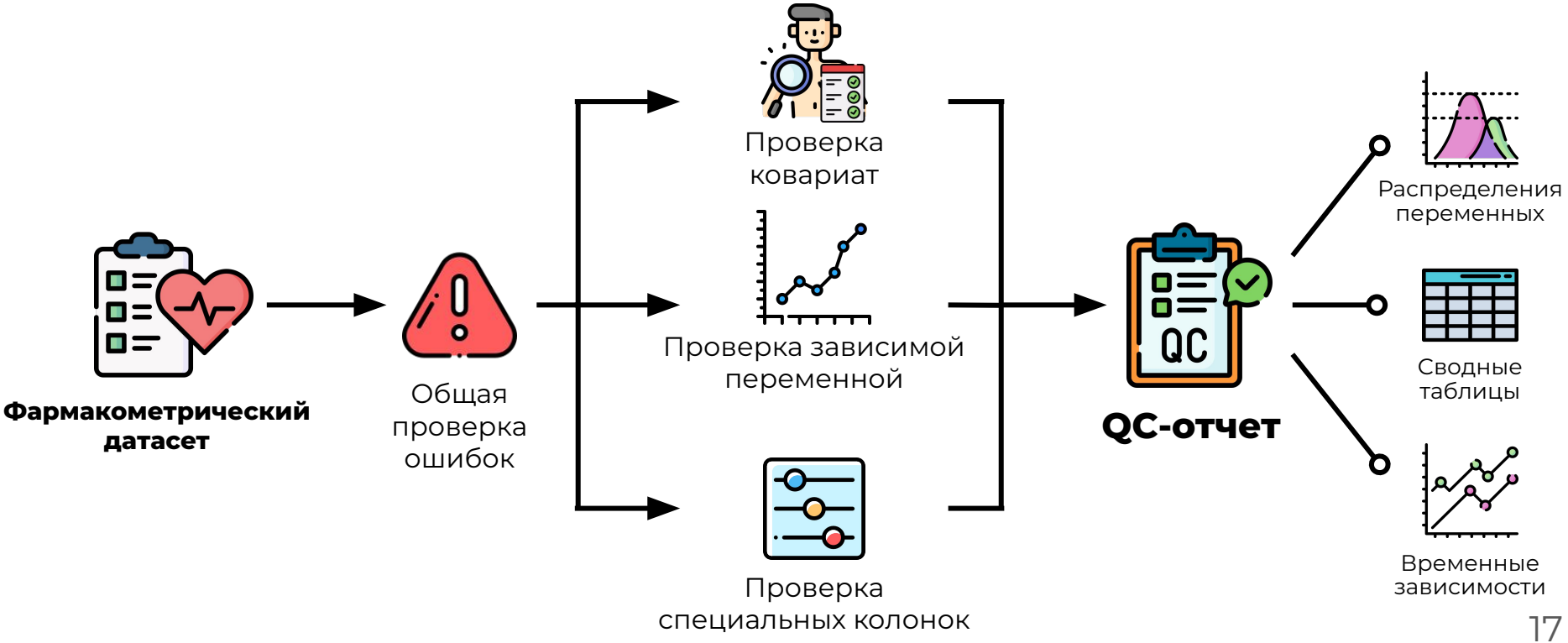
PHANDORIN

PHArmacometrics Nlme Datasets Observation, Review and INspection

PHANDORIN - инструмент на R для контроля качества фармакометрических датасетов



PHANDORIN


Фармакометрический датасет
Общая проверка ошибок
Проверка ковариат
Проверка зависимой переменной
Проверка специальных колонок
QC
QC-отчет
Распределения переменных
Сводные таблицы
Временные зависимости

PHANDORIN: формат ввода

```
run_qc(file_path = 'data/warfarin_data.csv',
       sep = ',',
       na_strings = c('NA', 'NaN', '.'),
       id_col = 'ID',
       dv_col = 'DV',
       time_col = 'TIME',
       continuous_covariates = c('AGE', 'WT'),
       categorical_covariates = c('SEX'),
       dvid_col = 'DVID',
       amt_col = 'AMT',
       dv_grouping_cols = c('DVID'),
       output_file = 'warfarin_report.html'
)
```

PHANDORIN: формат ввода

Обязательные параметры

Техническая информация:

- Где лежат данные?
- Как прочитать данные?
- Куда сохранить результат?

```
run_qc(file_path = 'data/warfarin_data.csv',
       sep = ',',
       na_strings = c('NA', 'NaN', '.'),
       id_col = 'ID',
       dv_col = 'DV',
       time_col = 'TIME',
       continuous_covariates = c('AGE', 'WT'),
       categorical_covariates = c('SEX'),
       dvid_col = 'DVID',
       amt_col = 'AMT',
       dv_grouping_cols = c('DVID'),
       output_file = 'warfarin_report.html'
)
```

PHANDORIN: формат ввода

Обязательные параметры

Столбец с ID пациентов

Столбец с зависимой переменной

Столбец со временем

```
run_qc(file_path = 'data/warfarin_data.csv',
       sep = ',',
       na_strings = c('NA', 'NaN', '.'),
       id_col = 'ID',
       dv_col = 'DV',
       time_col = 'TIME',
       continuous_covariates = c('AGE', 'WT'),
       categorical_covariates = c('SEX'),
       dvid_col = 'DVID',
       amt_col = 'AMT',
       dv_grouping_cols = c('DVID'),
       output_file = 'warfarin_report.html'
)
```

PHANDORIN: формат ввода

Дополнительные параметры

Ковариаты

- Численные
- Категориальные

Специальные колонки

Группировка наблюдений

```
run_qc(file_path = 'data/warfarin_data.csv',
       sep = ',',
       na_strings = c('NA', 'NaN', '.'),
       id_col = 'ID',
       dv_col = 'DV',
       time_col = 'TIME',
       continuous_covariates = c('AGE', 'WT'),
       categorical_covariates = c('SEX'),
       dvid_col = 'DVID',
       amt_col = 'AMT',
       dv_grouping_cols = c('DVID'),
       output_file = 'warfarin_report.html'
)
```

PHANDORIN: формат вывода

HTML-отчёт

PHANDORIN: формат вывода

Общая проверка ошибок

QC-вопрос 1

график-ответ

QC-вопрос 2

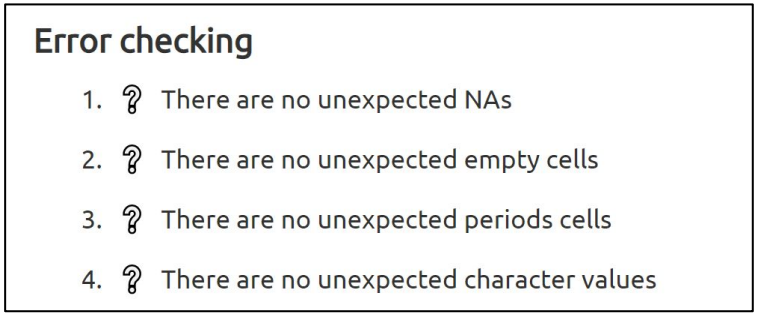
таблица-ответ

Проверка ковариат

QC-вопрос 3

график-ответ

Таблица-резюме



PHANDORIN: формат вывода

Общая проверка ошибок

QC-вопрос 1

график-ответ

QC-вопрос 2

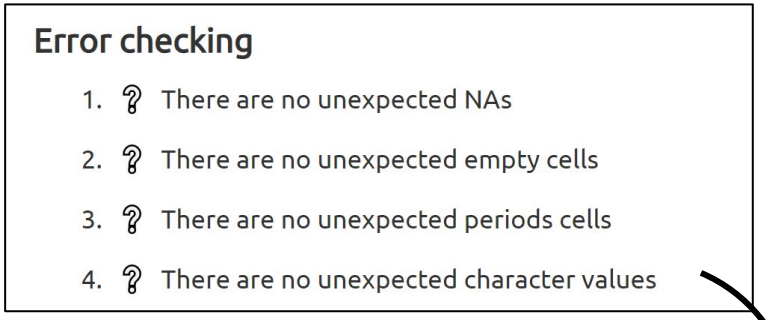
таблица-ответ

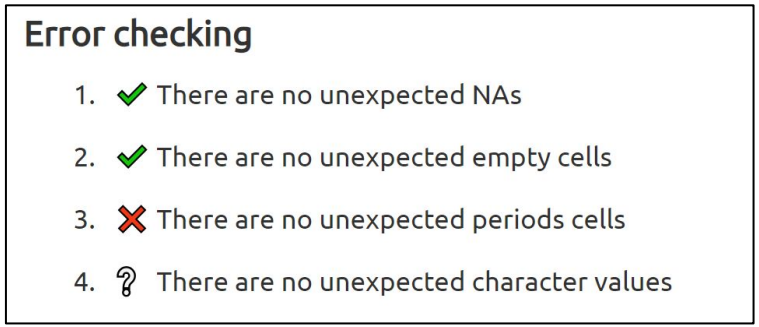
Проверка ковариат

QC-вопрос 3

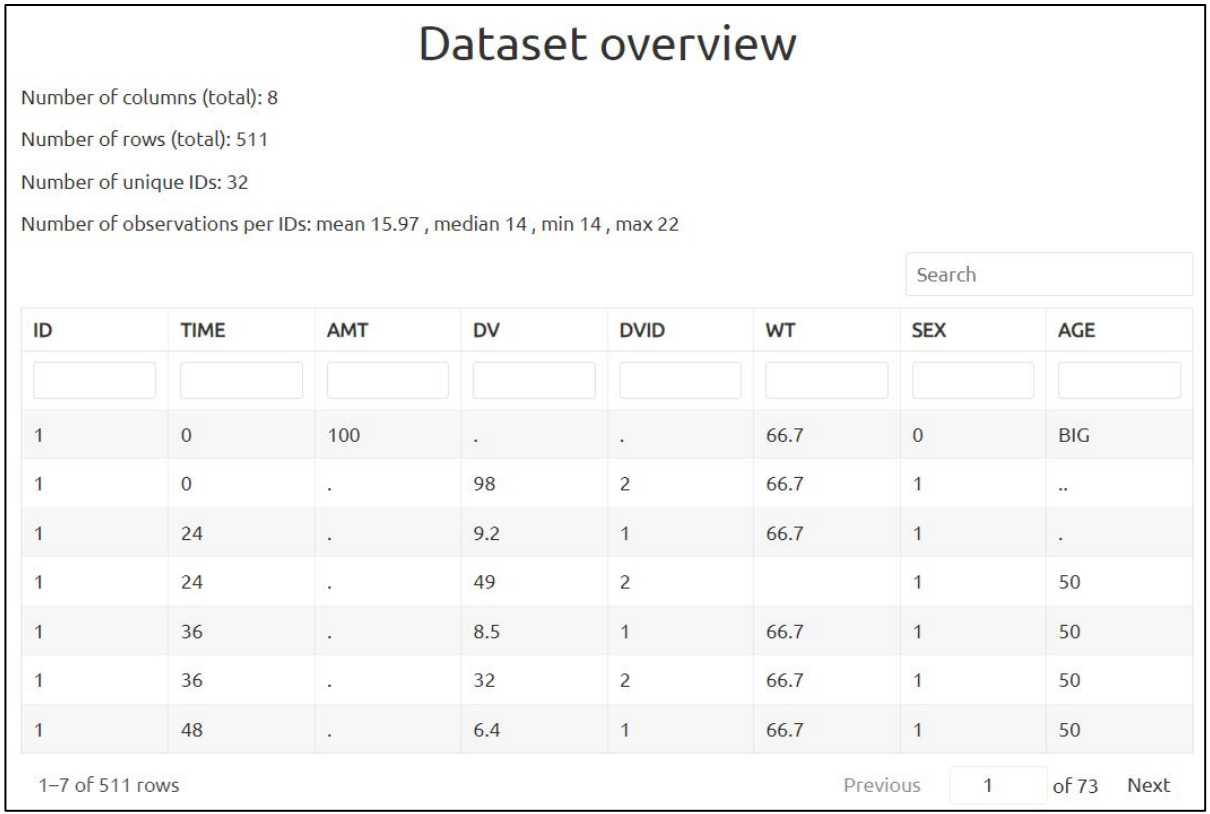
график-ответ

Таблица-резюме





Обзор данных



Dataset overview

Number of columns (total): 8

Number of rows (total): 511

Number of unique IDs: 32

Number of observations per IDs: mean 15.97 , median 14 , min 14 , max 22

Search

ID	TIME	AMT	DV	DVID	WT	SEX	AGE
1	0	100	.	.	66.7	0	BIG
1	0	.	98	2	66.7	1	..
1	24	.	9.2	1	66.7	1	.
1	24	.	49	2		1	50
1	36	.	8.5	1	66.7	1	50
1	36	.	32	2	66.7	1	50
1	48	.	6.4	1	66.7	1	50

1–7 of 511 rows Previous 1 of 73 Next

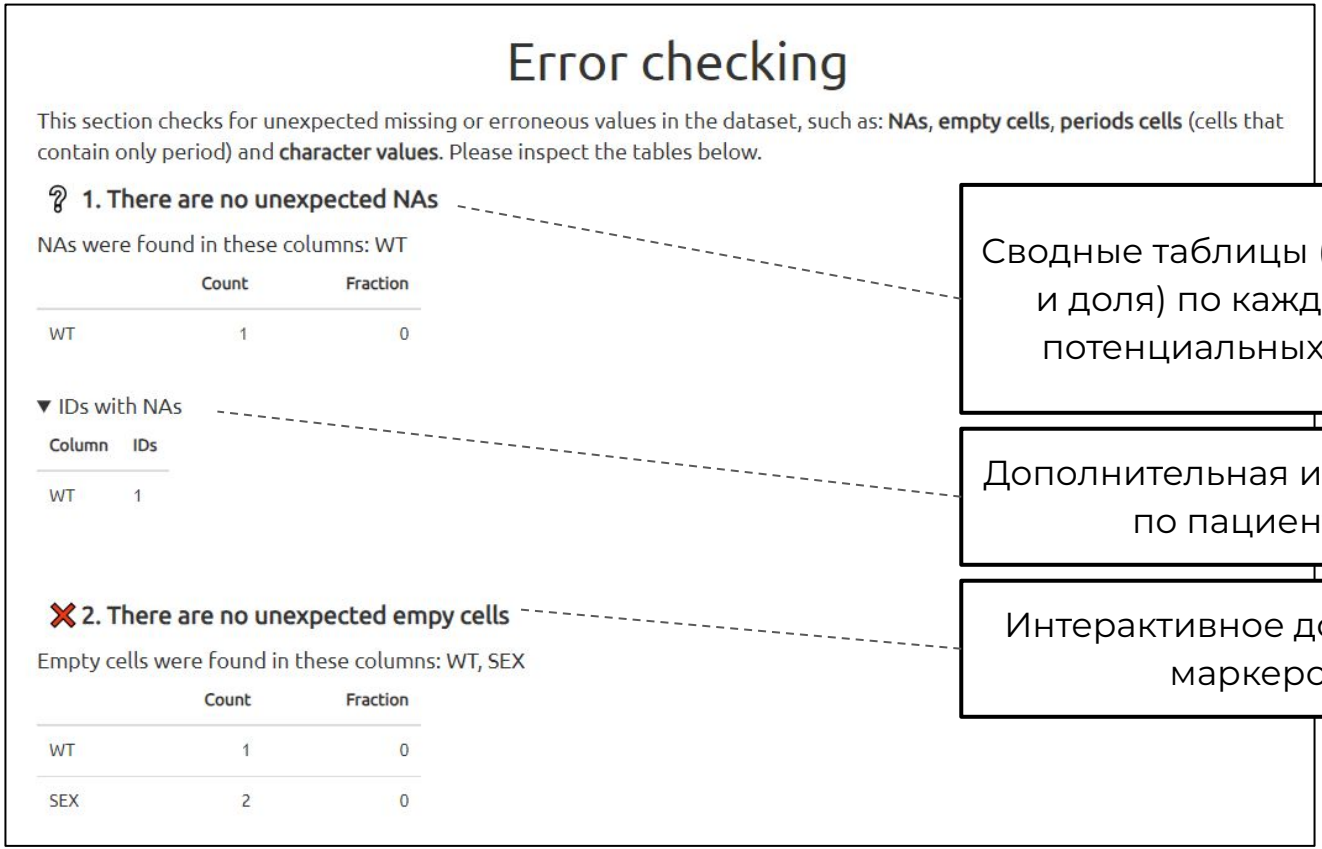
- Краткая статистика:
  - Кол-во записей
  - Кол-во пациентов
- Окно обзора данных:
  - Поиск по значениям
  - Фильтры по колонкам

Общая проверка ошибок

Имеются ли в наборе данных
пропущенные или ошибочные значения?

- **NA**
- **Пустые ячейки**
- **Ячейки с точкой**
- **Буквенные значения**

Общая проверка ошибок



Error checking

This section checks for unexpected missing or erroneous values in the dataset, such as: **NAs**, **empty cells**, **periods cells** (cells that contain only period) and **character values**. Please inspect the tables below.

1. There are no unexpected NAs

NAs were found in these columns: WT

	Count	Fraction
WT	1	0

▼ IDs with NAs

Column	IDs
WT	1

2. There are no unexpected empy cells

Empty cells were found in these columns: WT, SEX

	Count	Fraction
WT	1	0
SEX	2	0

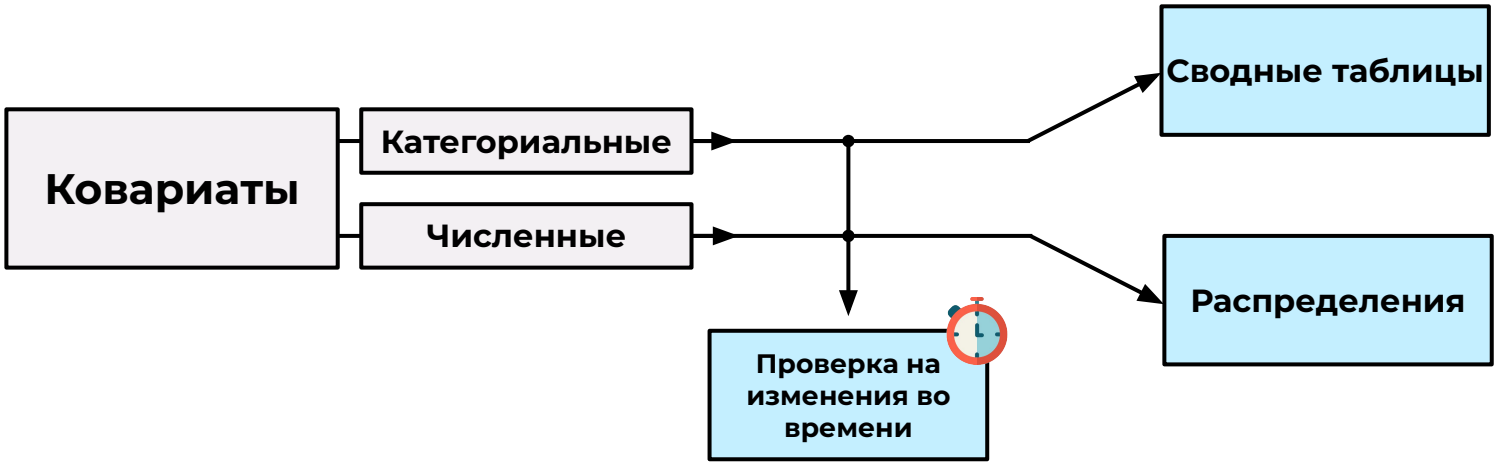
Сводные таблицы (количество и доля) по каждому типу потенциальных ошибок

Дополнительная информация по пациентам

Интерактивное добавление маркеров

Проверка ковариат

- Все ли изменяющиеся во времени ковариаты могут изменяться во времени?
- Распределения численных ковариат соответствуют ожиданиям?
- Значения категориальных ковариат соответствуют ожиданиям?

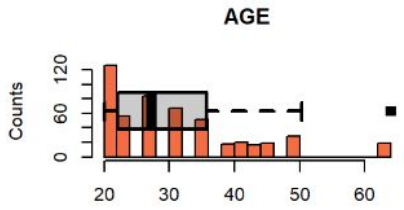


Проверка ковариат: распределения

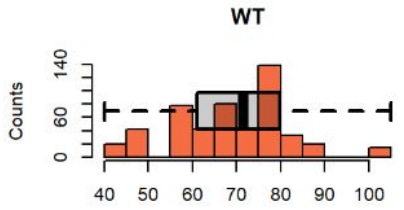
Numerical covariates

Covariates specified as continuous: *AGE* , *WT*

✔ **10. All numerical covariates have the expected distributions**



Mean: 31.72 SD: 10.53
Median: 28 Outliers 4 %
Min: 21 Max: 63
1st Qu.: 23 3rd Qu.: 36



Mean: 69.44 SD: 12.91
Median: 70 Outliers 0 %
Min: 40 Max: 102
1st Qu.: 60 3rd Qu.: 78

Categorial covariates

Covariates specified as categorical: *SEX*

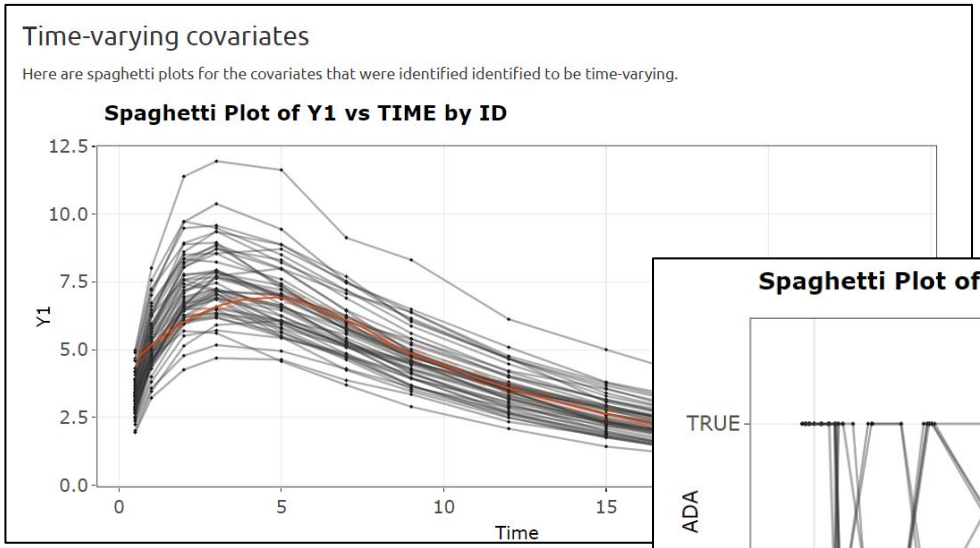
✘ **9. All categorical covariates have the expected levels**

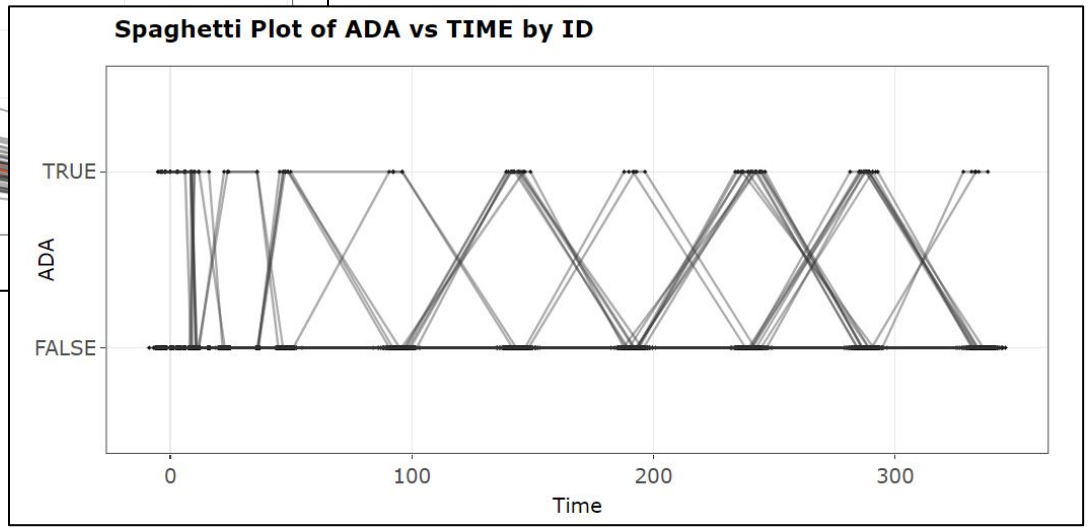
	IDs with value		Records per ID	
SEX	N	fraction	N	fraction
0	6	0.19	16.33	0.85
1	27	0.84	15.22	0.99
NA	1	0.03	2.00	0.10

Проверка ковариат: вариабельность

Time-varying covariates

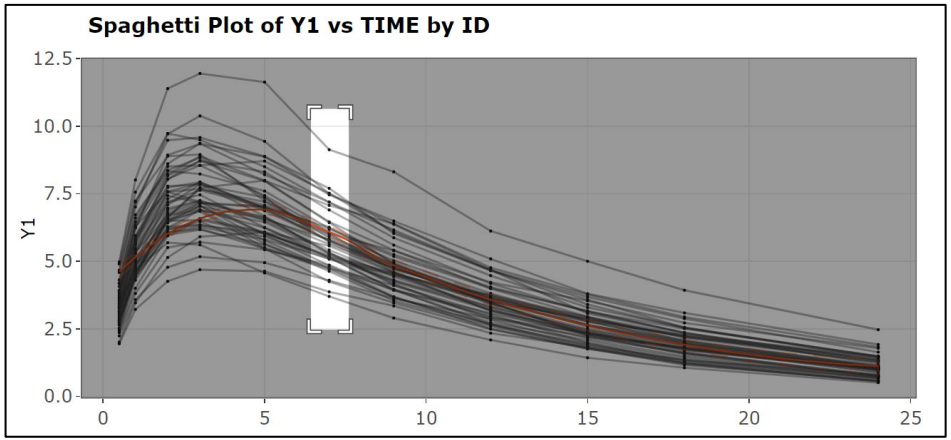
Here are spaghetti plots for the covariates that were identified identified to be time-varying.



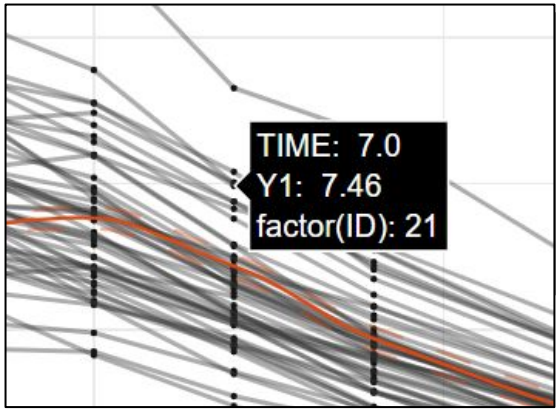


Все графики интерактивные

Можно приблизить определенный участок

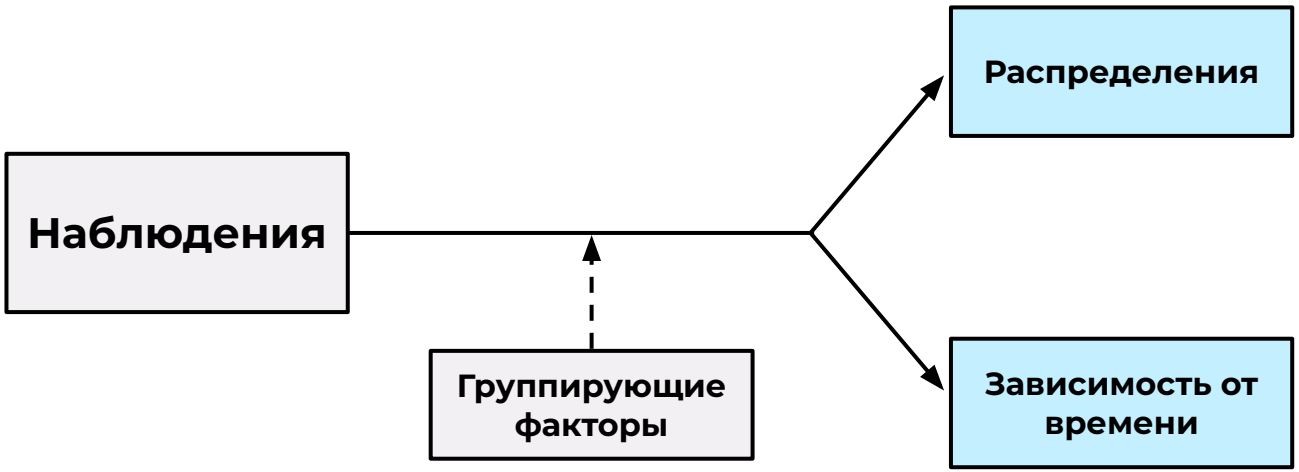


Можно выбрать определенные точки



Проверка зависимой переменной

- Есть не невозможные значения среди наблюдений?
- Наблюдаемые значения изменяются со временем согласно нашим ожиданиям?



Пример: DVID, STUDY

Проверка зависимой переменной

Dependent variable

✔ 11. Dependent variable have the expected distribution

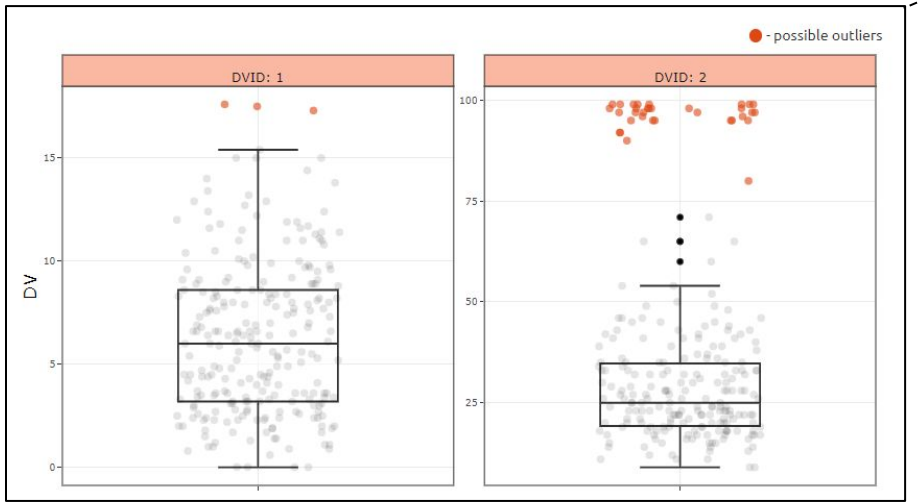
Dependent variable statistics (by provided DV-grouping indicators):

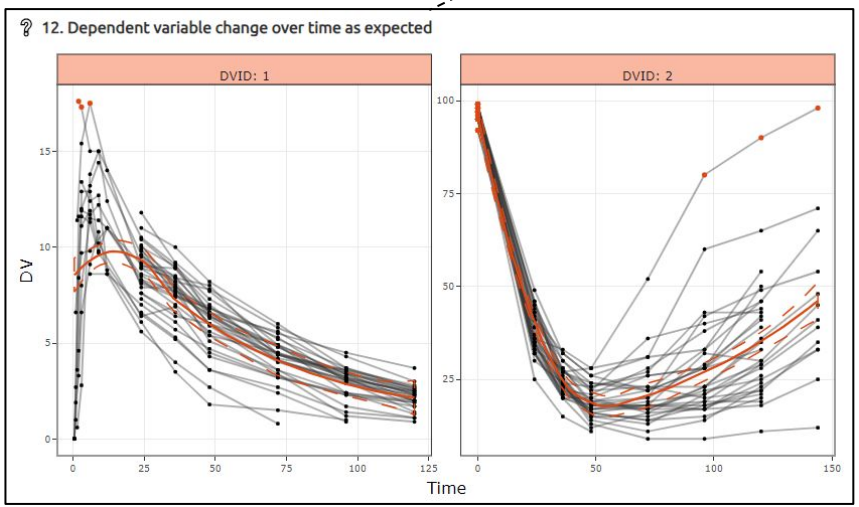
DVID	Mean	Median	Min	Max	SD	Q0.25	Q0.75
1	6.31	6	0.01	17.6	3.77	3.2	8.7
2	37.60	28	9.00	99.0	26.19	20.0	42.0

Общая статистика

Распределения

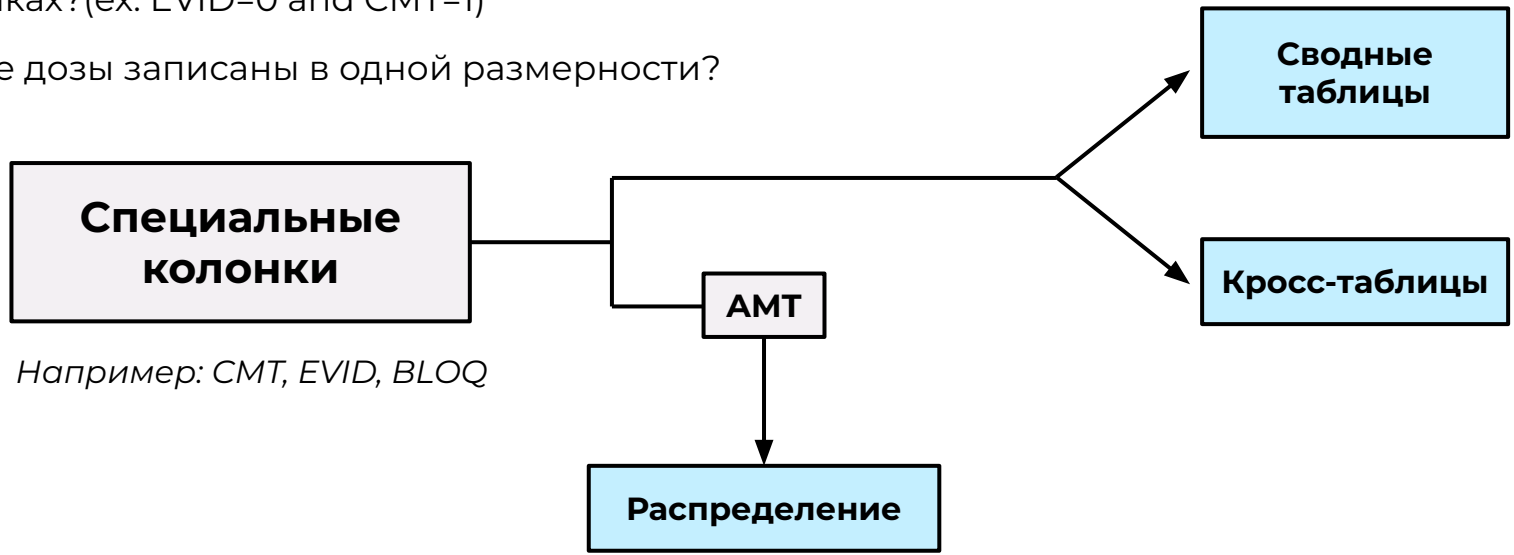
Временная зависимость





Проверка специальных колонок

- Есть ли неожиданные значения в специальных колонках?
- Есть ли неожиданные комбинации значений в специальных колонках?(ex. EVID=0 and CMT=1)
- Все ле дозы записаны в одной размерности?



Проверка специальных колонок

Special columns

Special columns specified by user: *DVID* , *AMT*

✔ 13. All special columns have the expected values and counts

	IDs with value		Records per ID	
DVID	N	fraction	N	fraction
1	32	1	7.72	0.48
2	32	1	7.25	0.46
NA	32	1	1.00	0.06

Статистика по колонкам

AMT information

Summary | Detailed

	IDs with value		Records per ID	
AMT	N	fraction	N	fraction
AMT = 0 or "."	32	1	14.97	0.94
AMT > 0	32	1	1.00	0.06

Отдельная информация про дозы

▼ Cross-tables for NONMEM-specific columns

	DVID = 1	DVID = 2	DVID = NA
AMT = 0 or "."	247	232	0
AMT > 0	0	0	32

Кросс-таблицы

PHANDORIN: формат вывода

Общая проверка ошибок

QC-вопрос 1

график-ответ

QC-вопрос 2

таблица-ответ

Проверка ковариат

QC-вопрос 3

график-ответ

Таблица-резюме

Error checking

1. ✔ There are no unexpected NAs
2. ✔ There are no unexpected empty cells
3. ✔ There are no unexpected periods cells
4. ✘ There are no unexpected character values

Dates and times

5. ✔ All dates are in the expected formats

Covariates

6. ✔ All time-invariant covariates are expected to be time-invariant
7. ✔ All time-varying covariates are expected to be time-varying
8. ✔ All time-varying covariates change over time as expected
9. ✔ All categorical covariates have the expected levels
10. ✔ All numerical covariates have the expected distributions

Dependent variable

11. ✔ Dependent variable have the expected distribution
12. ✘ Dependent variable change over time as expected

NONMEM-specific columns

13. ✔ All special columns have the expected values and counts
14. ✔ There are no unexpected dosages and all dosages are in the consistent dimensionality
15. ✔ There are no records with impossible combination of special columns values

Заключение

Разработан инструмент PHANDORIN для автоматического и унифицированного контроля качества фармакометрических данных

Инструмент был отлажен на 51 датасете (публичные датасеты Monolix и nlmixer, датасеты M&S Decision).

На данный момент PHANDORIN нацелен на QC-анализ **PKPD** данных.

Будущие планы:

- Внедрить дополнительные проверки (необходимые **вам)**;
- Расширить возможности PHANDORIN на другие типы данных (time-to-event данные, цензурированные данные, ...);

Спасибо за внимание!