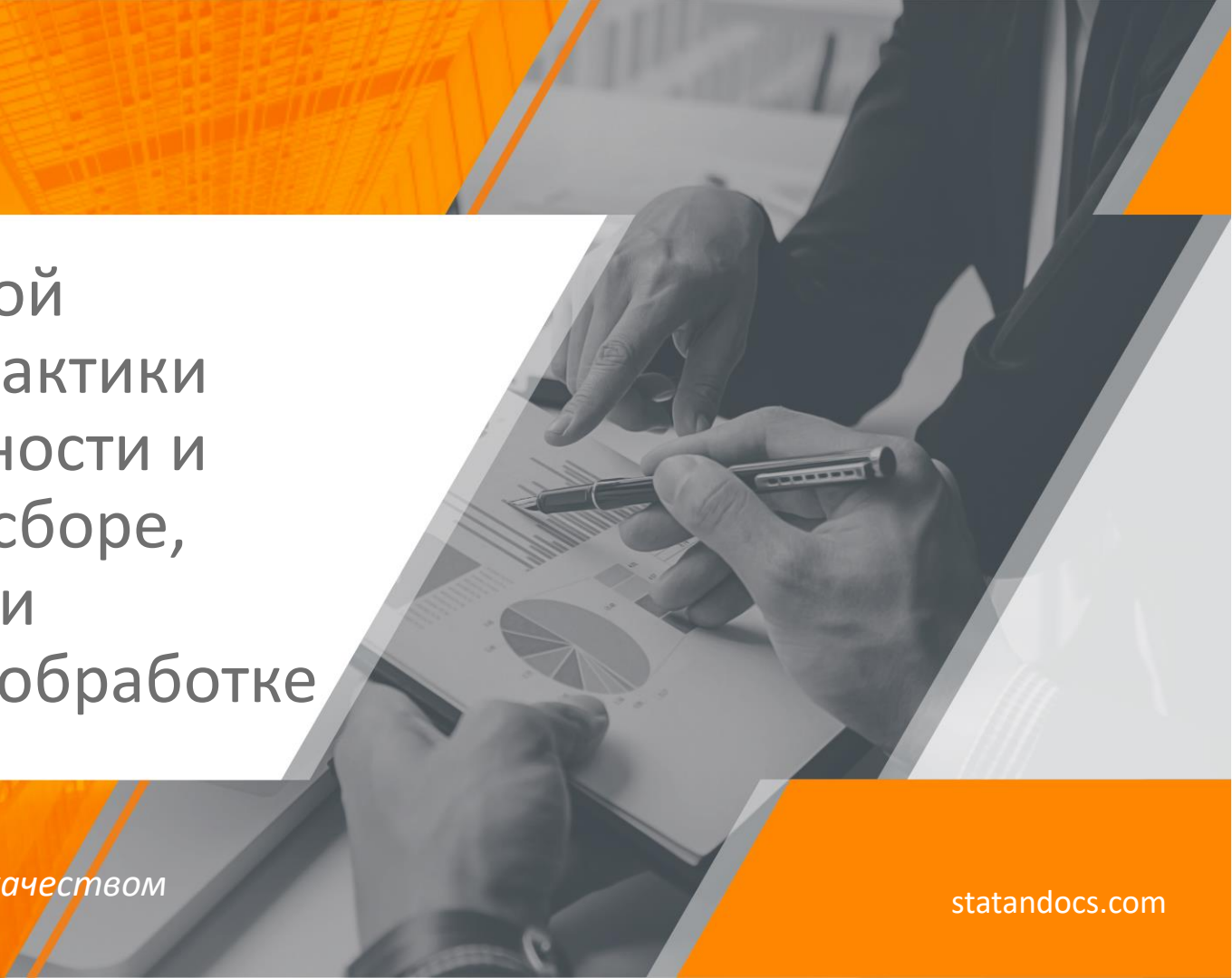
Данные реальной клинической практики (RWD): возможности и сложности при сборе, интерпретации и статистической обработке



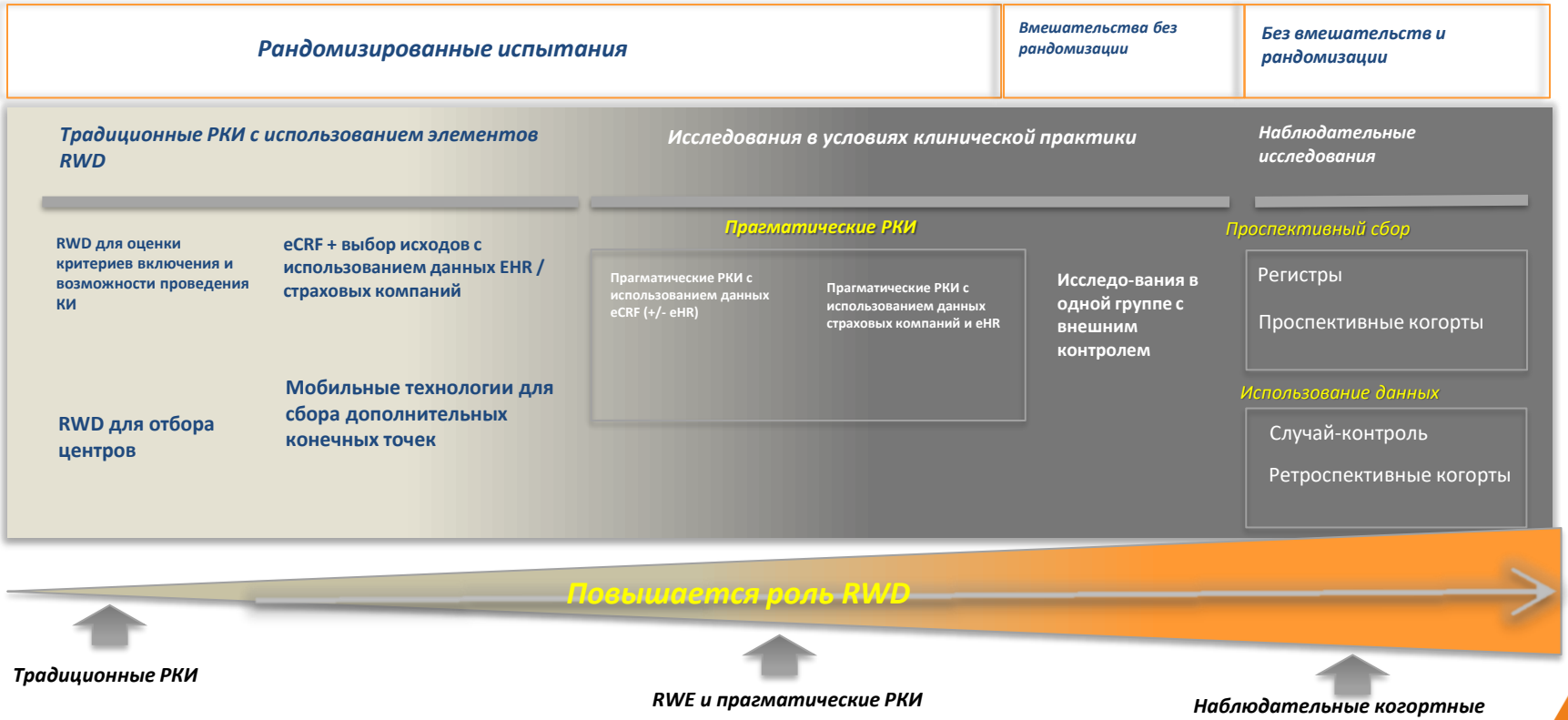
Александр Солодовников
Директор по управлению качеством
ООО «Статэндокс»

statandocs.com

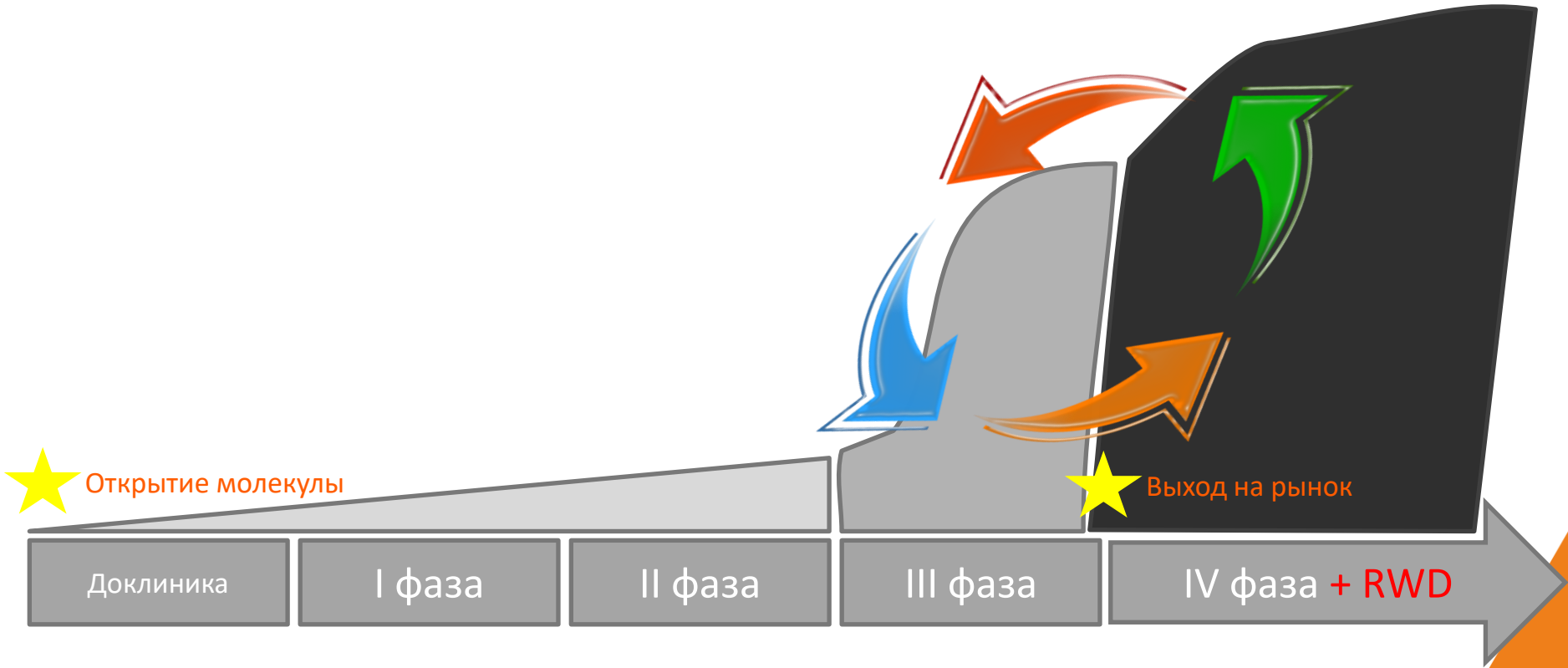
Определение RWD и RWE

- "**данные реальной клинической практики**" - данные, относящиеся к состоянию здоровья пациента и (или) к процессу оказания медицинской помощи, ***полученные из различных источников***.
- "**доказательства, полученные на основе данных реальной клинической практики**" - клинические доказательства в отношении применения и потенциальной пользы или риска применения лекарственного препарата, полученные на основе сбора и анализа данных реальной клинической практики

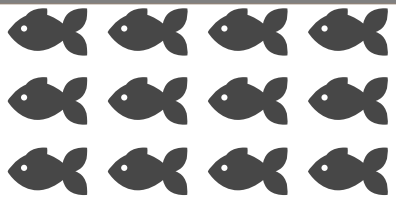
RWD в разных видах исследований



Сравнительный объем данных



Клинические исследования vs RWE



КИ

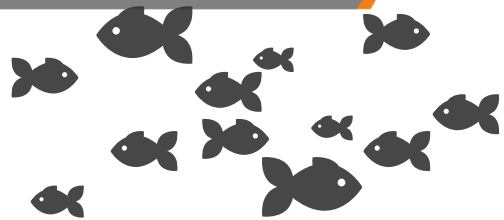
Проспективные

Интервенционные (фиксированный режим лечения)

Рандомизированные

Контрольная группа и группа лечения

Гомогенная селективная группа



RWE

Проспективные, одномоментные, ретроспективные

Наблюдательные (фактический гибкий режим лечения)

Распределение по группам не контролируется

Контрольная группа формируется в естественных условиях (либо отсутствует)

Гетерогенная стихийно формирующаяся группа

В КИ мы узнаем ‘может ли препарат действовать’,
а в RWE мы изучаем ‘действует ли препарат’

Источники получения RWD

Исследования

- Клинические
- Прагматические
- Фармако-экономические
- Фармако-эпидемиологические
- Наблюдательные

Система здравоохранения

- ЛПУ
- Страховые компании
- Лаборатории
- Регистры (диагнозов, пациентов)
- Аптеки

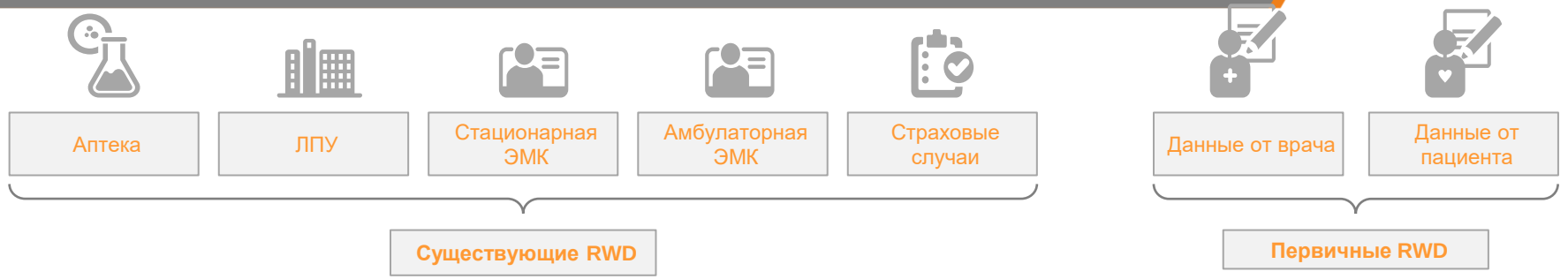
Пациент / человек

- Социальные сети
- Форумы
- Социальные СМИ
- Приложения для ввода информации о здоровье

Медицинские и носимые устройства

- Системы непрерывного мониторинга
- Глюкометры
- Тонометры
- Фитнес-браслеты, смарт-часы и т.д.

Существующие vs первичные RWD



Существующие данные позволяют:

- Охарактеризовать паттерны лечения и расходов
- Предоставить контекст для исследований в одной группе и оценки сигналов безопасности
- Сравнить действенность препаратов

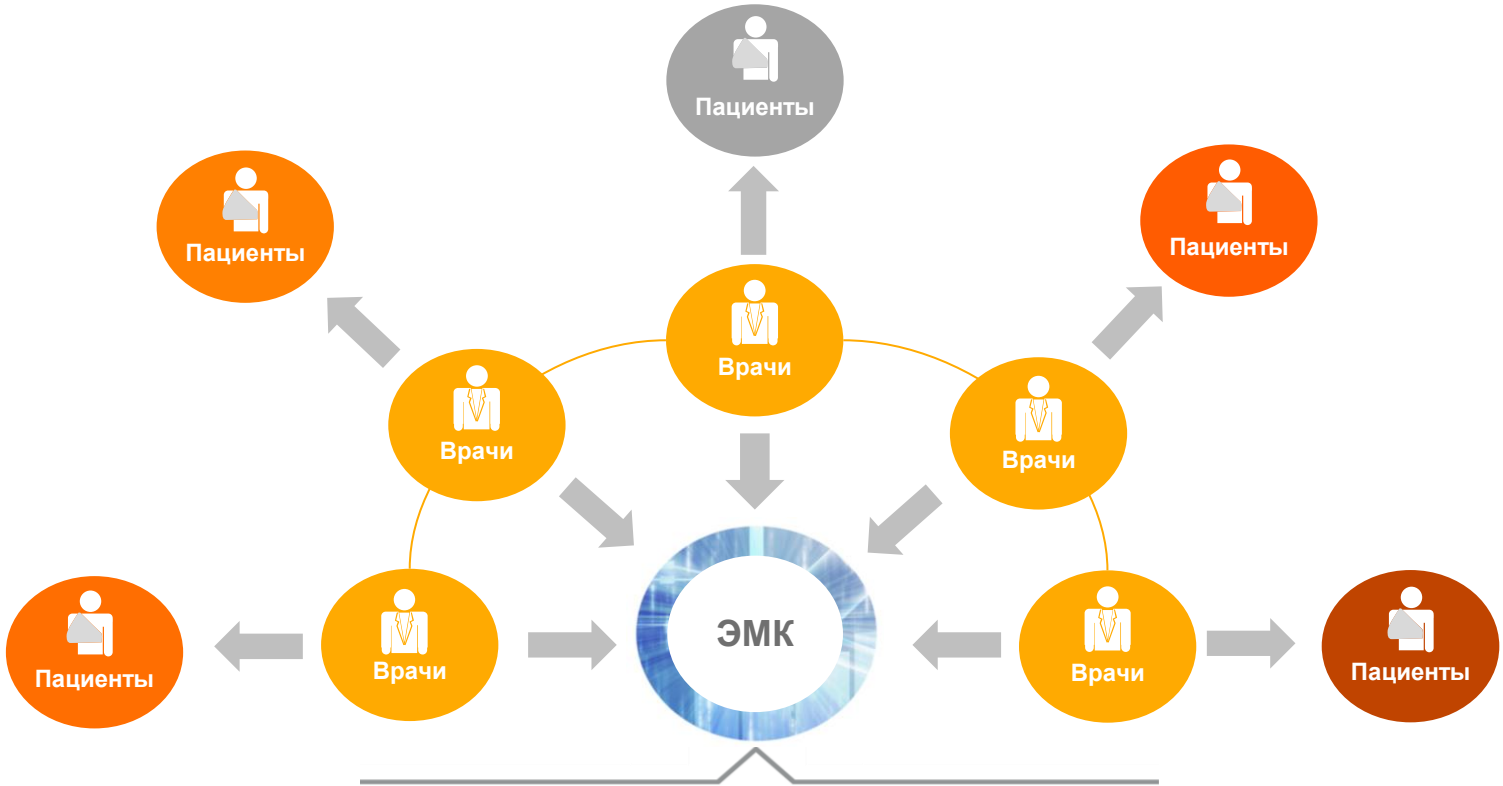
Первичные данные позляют:

- Оценить системный клинический вклад
- Изучить исходы с точки зрения врача
- Изучить исходы с точки зрения пациента

Dreyer NA, Rodriguez AM. Fast route to evidence development for value in health care. *Curr Med Res Opin.* 2016 Jul;11:1-4.

Сложность в многообразии…

Электронная медицинская карта (ЭМК)



stat&docs

СОБИРАЕМ RWD

ДАННЫЕ, МЕТАДАННЫЕ, КОНТЕКСТ

statandocs.com

Требования к процессу получения RWD

получение **надежных (валидных)**, **точных**, ***максимально полных результатов*** при **соблюдении воспроизводимой и прозрачной процедуры их получения**

Многослойность данных

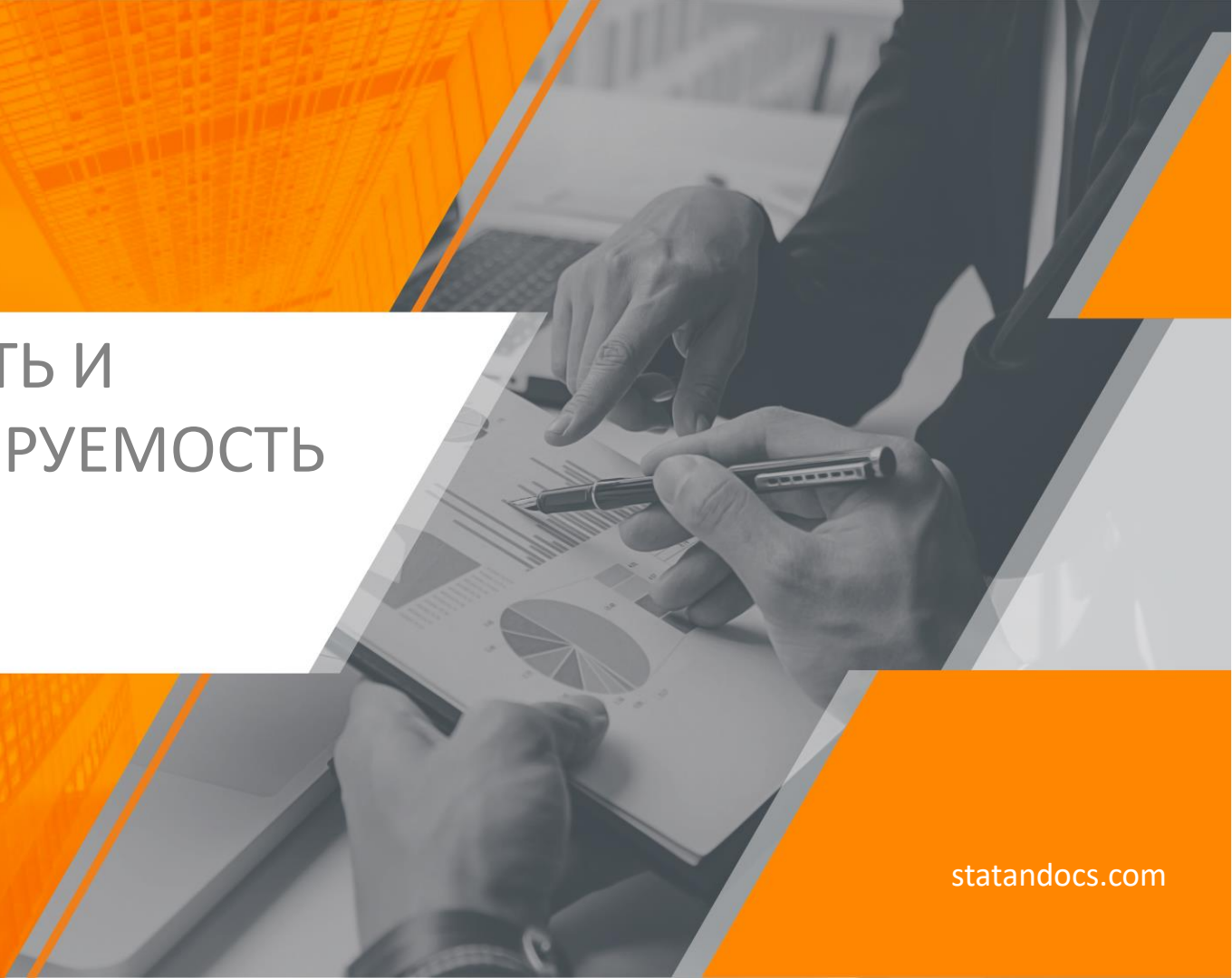
Ключевые моменты подходов к сбору и обработке RWD

- оценка *особенностей* источников данных
- оценка *наличия данных и причин отсутствия*
- оценка *множественности* источников
- анализ целостности и уникальности *(идентифицируемости)* данных
- *возможность, целесообразность и методология восполнения* пропущенных данных
- *анализ чувствительности*
- **интерпретируемость результатов относительно поставленных целей и задач**

Особенности RWD

- Отсутствие исходной нацеленности RWD на решение конкретных научных задач
- Варианты структуры:
  - структурированные
  - неструктурированные
- Пропущенные (missing data) и несогласованные (inconsistent data) данные

УНИКАЛЬНОСТЬ И ИДЕНТИФИЦИРУЕМОСТЬ ДАННЫХ



statandocs.com

Уникальность данных

- Проблема множественности источников:
  - Электронная история болезни
  - База данных лаборатории
  - Записи/данные от пациента
  - Регистры и прочие системы первичного сбора информации (включая EDC)

Все они могут содержать одну и ту же информацию!

Идентифицируемость данных

- Наличие единого идентификатора в разных источниках
- Использование разных идентификаторов и их «конвертируемость» друг в друга
- Валидность / уникальность идентификатора
- **Полнота прочей мета-информации**

СОГЛАСОВАННОСТЬ ДАННЫХ

statandocs.com

Согласованность данных

- Варианты несогласованности:
  - Разный объем информации в источниках
  - Декларируемое отсутствие информации в одном источнике при наличии информации в другом источнике (источниках)
  - Противоречие данных в разных источниках
  - Противоречие мета-информации в разных источниках

Согласованность данных

- Борьба с несогласованностью:
  - Присвоение «категорий валидности» источникам и отдельным элементам данных
  - Применение арбитража (внешний контрольный источник)
  - Точечные запросы (при наличии технической возможности)
  - Кросс-проверки по связанным элементам данных

НЕПОЛНОТА ДАННЫХ (MISSIGNESS)

Неполнота данных (missingness)

- Неполнота данных (Missingness)—существование пропущенных данных **И механизма (причины), который объясняет причину отсутствия данных**
- Механизм отсутствия данных:
  - MCAR – missing completely at random
  - MAR – missing at random
  - **MNAR – missing not at random**
- Доля пропущенных данных – прямо связана с качеством статистических выводов
- Уровни отсутствия данных
  - Единица наблюдения (пациент, участник)
  - Параметр измерения
- Паттерн отсутствующих данных
  - Однофакторный
  - Монотонный
  - Произвольный

Неполнота данных (missingness)

- Статистические методы восполнения:
  - Прямая импутация (LOCF, BOCF),
  - Смешанная модель повторных измерений (MMRM),
  - Множественная импутация (MI),
  - Присвоение «весов» отсутствующим измерениям (weighting),
- Допущения и паттерны неполноты данных для определения статистических методов:
  - MCAR, MAR, MNAR
  - Допущения для аналитических моделей

Механизм отсутствия данных

- Missing Completely at Random (MCAR) – полностью случайное отсутствие данных
  - Отсутствие или наличие данных никак не связано с ненаблюдаемым результатом
- Missing at Random (MAR) – случайное отсутствие данных
  - **Сама по себе неполнота данных неслучайна**. **Однако** при этом неполнота условно случайна и не зависит от ненаблюдаемого результата
- Missing Not at Random (MNAR) – неслучайное (систематическое) отсутствие данных
  - Любые другие условия (*«неигнорируемый неответ»*)
  - Отсутствие данных зависит от значения ненаблюдаемого результата

Основные методы восполнения данных

- Удаление
  - Попарное удаление
  - Удаление случая
- Импутация
  - “Простая” импутация (среднее, медиана, худшее наблюдение, последнее наблюдение и т.д.)
    - Этот метод предполагает, что в реальности известно больше информации, по сравнению с тем, что доступно на момент анализа, и восполненные значения рассматриваются как реальные данные (это приводит к искусственному занижению случайной ошибки и возможному искажению значений p)
  - Частичная импутация
    - В случае использования EM-алгоритма (Expectation-Maximization) полученная оценка предполагает наличие полных данных, но при этом учитывает паттерн неполноты
  - Множественная импутация и максимальное правдоподобие
    - В целом считается предпочтительной по сравнению с другими методами, поскольку позволяет контролировать как случайную ошибку, так и корректно интерпретировать уровень значимости
    - Требует проведения математического моделирования

Примеры реализации импутации при MNAR

stat&docs

Article

SAGE Open
January-March 2018: 1–12

DOI: 10.1177/2158244018757584
journals.sagepub.com/home/sgo
SAGE

Multiple Imputation for Dichotomous MNAR Items Using Recursive Structural Equation Modeling With Rasch Measures as Predictors

Celeste Combrinck[1], Vanessa Scherman[2], David Maree[1], and Sarah Howie[1]

Abstract

Missing Not at Random (MNAR) data present challenges for the social sciences, especially when combined with Missing Completely at Random (MCAR) data for dichotomous test items. Missing data on a Grade 8 Science test for one school out of seven could not be excluded as the MNAR data were required for tracking learning progression onto the next grade. Multiple imputation (MI) was identified as a solution, and the missingness patterns were modeled with IBM Amos applying recursive structural equation modeling (SEM) for 358 cases. Rasch person measures were utilized as predictors. The final imputations were done in SPSS with logistic regression MI. Diagnostic checks of the imputations showed that the structure of the data had been maintained, and that differences between MNAR and non-MNAR missing data had been accounted for in the imputation process.

Примеры реализации импутации при MNAR и MAR

stat&docs

npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Check for updates

Imputation of missing values for electronic health record laboratory data

Jiang Li[1], Xiaowei S. Yan[2], Durgesh Chaudhary[1], Venkatesh Avula[1], Satish Mudiganti[2], Hannah Husby[2], Shima Shahjouei[1], Ardavan Afshar[3,7], Walter F. Stewart[4], Mohammed Yeasin[5], Ramin Zand[1] and Vida Abedi[1,6]✉

Laboratory data from Electronic Health Records (EHR) are often used in prediction models where estimation bias and model performance from missingness can be mitigated using imputation methods. We demonstrate the utility of imputation in two real-world EHR-derived cohorts of ischemic stroke from Geisinger and of heart failure from Sutter Health to: (1) characterize the patterns of missingness in laboratory variables; (2) simulate two missing mechanisms, arbitrary and monotone; (3) compare cross-sectional and multi-level multivariate missing imputation algorithms applied to laboratory data; (4) assess whether incorporation of latent information, derived from comorbidity data, can improve the performance of the algorithms. The latter was based on a case study of hemoglobin A1c under a univariate missing imputation framework. Overall, the pattern of missingness in EHR laboratory variables was *not at random* and was highly associated with patients' comorbidity data; and the multi-level imputation algorithm showed smaller imputation error than the cross-sectional method.

npj Digital Medicine (2021)4:147 ; https://doi.org/10.1038/s41746-021-00518-0

Примеры реализации импутации при MAR

ОСОБЕННОСТИ РАБОТЫ С ДАННЫМИ, ХАРАКТЕРИЗУЮЩИМИ ЗДОРОВЬЕ НАСЕЛЕНИЯ: ЗАПОЛНЕНИЕ ПРОПУСКОВ В ДАННЫХ

30.03.2020 г.

DOI: 10.21045/2071-5021-2020-66-1-12

Аладышкина А.С., Лакшина В.В., Леонова Л.А., Максимов А.Г.
Национальный исследовательский университет «Высшая школа экономики», Нижний Новгород, Россия

Резюме

Актуальность. В ряде случаев базы данных показателей, использующихся как для исследований в области здравоохранения, так и для решения различных социально-экономических задач, содержат существенное количество пропущенных значений. Для повышения эффективности работы с такими базами возникает необходимость заполнения пропусков в данных. Эта необходимость обусловлена потерей информации при простом игнорировании пропущенных значений и возможностью получения в этом случае смещенных и несостоятельных результатов.

Цель исследования: оценить применимость алгоритма MICE (multivariate imputation by chained equations) для восстановления пропущенных значений в приложении к данным, релевантным задаче исследования зависимости предложения труда от показателей здоровья населения.

Материал и методы. Исследование проводилось на основе данных RLMS HSE. Для восстановления пропущенных значений был применен алгоритм MICE, основанный на использовании метода Монте-Карло по схеме марковской цепи для получения апостериорных распределений переменных, содержащих пропуски.

Результаты. Проведенный анализ показал наличие существенной доли пропусков в значениях выбранных переменных, включающих в себя показатели здоровья и социально-экономические характеристики респондентов. Произведено восстановление пропущенных значений переменных алгоритмом MICE, результаты работы алгоритма проверены на сходимость. Получены эмпирические оценки плотностей и функций вероятности для восстановленных данных. В качестве примера восстановленные данные применены для оценки параметров пространственной панельной регрессии, для каждого параметра по правилу Рубина рассчитаны стандартные ошибки с учетом проведенной импутации, а также доля дисперсии из-за пропусков в данных.

Справочная таблицы по методам восполнения данных

Метод	Описание	Область применения
Multiple imputation by chained equations (MICE)	Импутация проводится путем регрессии по всем показателям; значение считается MAR	Восполнение дихотомических и количественных показателей
Amelia	Метод множественной импутации, основанный на общем бутстрепе по отсутствующим данным	Может использоваться только для нормально распределенных количественных показателей
MissForest	Алгоритм случайных подпространств; непараметрический метод импутации; создает модель «случайный лес» на основе фактически наблюдаемых значений	Восполнение качественных и количественных показателей
Harrell Miscellaneous (Hmisc)	Множество вариантов импутации, включая методы на основе среднего и мин/макс методы	Восполнение дихотомических и количественных показателей
Multiple Imputation (mi)	Использование Байесовской регрессии; выявляет и корректирует коллинеарность между переменными, а также величину случайной ошибки	Восполнение качественных и количественных показателей
Distributional Based Imputation (DBI)	Использует однофакторный подход и создает распределение значений на основе ожидаемого среднего и стандартного отклонения имеющихся значений; замена значений происходит случайным методом, таким образом, в модели сохраняется случайная ошибка	Восполнение только количественных показателей; результаты лучше, чем у моделей на основе среднего или медианы

Примеры использования RWD в регистрационном досье

Drug	Indication	Sponsor Year	Type of RWD submitted as historical control	Endpoint for comparative efficacy
Blincyto[1]	Sub-type of acute lymphoblastic leukemia (ALL)	Amgen 2018	Medical records for 121 patients over 8 years from 14 institutions in the US, Canada, Australia - Prospectively planned, retrospective study	CR
Brineura[2]	Batten disease (CLN2)	BioMarin 2017	Disease registry of 69 children (42 included): records & patient interviews - Prospectively planned, mostly retrospective study	CLN2 rating scale (motor, language)
Bavencio[3]	Metastatic Merkel cell carcinoma	EMD Serono 2017	Electronic medical records from 686 patients (14 included) from community and academic centers - Prospectively planned, retrospective study	RECIST
Exondys 51[4]	Duchenne Muscular Distrophy	Sarepta 2016	2 natural disease history cohorts (Belgium & Italy) of about 90 patients each (13 included) - Post-hoc retrospective study	6-min walking test

Примеры использования RWD в регистрационном досье

- Все 4 примера на предыдущем слайде:
  - **Тщательный протокол отбора популяции** (независимые оценщики, система независимого арбитража спорных случаев), что позволило оптимизировать объем выборки сравнения
  - ***Используемые конечные точки имели низкий % неполноты данных***
  - **Влияние неполноты** данных оценивалось **при помощи расширенного анализа чувствительности** *(в одном случае с контролем в виде проспективного сбора данных)*

СПАСИБО ЗА ВНИМАНИЕ!

statandocs.com