Power BI как вспомогательный инструмент биостатистика

Дмитриева Н. Ю.
ЗАО «Астон Консалтинг», Москва, Россия

Аннотация

Визуализация результатов исследований на основе данных реальной клинической практики (RWD) является обязательной при проведении презентаций и публикации отчётов. Одним из дополнительных инструментов, позволяющих настраивать сложные отчёты, является сервис бизнес-аналитики от Microsoft Power BI. В статье рассмотрены различные способы загрузки, обработки и визуализации данных клинических регистров с помощью указанного программного обеспечения.

Ключевые слова: Power BI; Python; R; клинические регистры; наблюдательные программы; данные реальной клинической практики; RWD; RWE

Для цитирования: Дмитриева Н. Ю. Power BI как вспомогательный инструмент биостатистика // *Реальная клиническая практика: данные и доказательства*. 2022;1(2):36-39. https://doi.org/10.37489/2782-3784-myrwd-10.

Поступила: 01 марта 2022 г. **Одобрена**: 04 марта 2022 г. **Опубликована**: 15 марта 2022 г.

Power BI as an auxiliary tool for biostatistics

Dmitrieva N. Yu.
ZAO Aston Consulting, Moscow, Russia

Abstract

Visualization of research results based on data from real clinical practice (RWD) is mandatory when conducting presentations and publishing reports. One of the additional tools that allow you to customize complex reports is the business intelligence service from Microsoft Power BI. The article discusses various ways of loading, processing and visualizing data from clinical registers using the specified software.

Keywords: Power BI; Python; R; clinical registries; observation programs; real-world data; RWD; RWE

For citation: Dmitrieva NYu. Power BI as an auxiliary tool for biostatistics. *Real-World Data & Evidence*. 2022;1(2): 36-39. https://doi.org/10.37489/2782-3784-myrwd-10.

Received: March 01, 2022. **Accepted:** March 04, 2022. **Published:** March 15, 2022.

Введение

Понятная и точная визуализация данных, полученных на основе реальной клинической практики (*англ.* real-world data; RWD), является одним из важных моментов при их анализе и в дальнейшем при презентации результатов исследования.

На наш взгляд, применение инструментов, используемых в бизнес-аналитике, может дать дополнительные возможности для построения интерактивных отчётов по данным регистров больных и нозологий, медицинских баз данных, электронных медицинских карт и т. д. [1].

В данной статье речь пойдёт об использовании такого инструмента от компании Microsoft, как Power BI, для анализа медицинских данных, в том числе клинических регистров [2]. Power BI представляет собой комплексное программное обеспечение для бизнес-аналитики, состоящее из нескольких самостоятельных продуктов, позволяющих создавать единую информационно-аналитическую систему обеспечения принятия решений на основе интегрированного анализа данных из различных источников для обеспечения максимальной глубины анализа и детализации данных [3].

Загрузка и обработка данных

Power BI позволяет подключиться к большому количеству различных источников данных. Осуществляется подключение с помощью редактора запросов (Power Query), выполняющего загрузку и очистку данных.

К основным возможностям Power Query относятся объединение и добавление данных, позволяющие комбинировать данные из нескольких источников данных, их фильтрация, сведение и развёртывание столбцов, а также добавление пользовательских вычисляемых столбцов. Для выражения всех подобных комбинаций данных используется язык формул Power Query M [4].

В нашем случае для ведения наблюдательных программ или программ лабораторной диагностики мы используем два решения — это универсальный программный комплекс для сбора, обработки и управления территориально распределёнными клинико-эпидемиологическими данными в режиме удалённого доступа Quinta, созданный на базе Microsoft Dynamics CRM 2011 [5], и самостоятельно разрабатываемые web-приложения.

Поддерживаемый Power BI открытый веб-протокол для запроса и обновления данных позволяет быстро взаимодействовать с CRM, используя в качестве запросов HTTP-команды, и получать необходимые данные в формате XML.

Также сбор данных и обработка могут быть осуществлены с помощью скриптов, написанных на Python или R.

Таким образом, мы можем при создании отчёта объединять различные источники данных, создавая необходимую схему данных.

Визуализация данных

После сбора необходимых данных мы переходим к процессу создания отчёта. Здесь проявляются все те преимущества, которые дают отчёты, построенные с помощью инструментов бизнес-аналитики:

- возможность расширенной аналитики;
- наглядность и визуализация;
- интерактивность;
- поиск инсайтов;
- доступ с любого устройства.

Таким образом, мы можем выстраивать гибкую систему фильтров и находить неявные взаимодействия между данными.

Также настраивать более гибкую визуализацию данных позволяют меры, с помощью которых можно производить дополнительные вычисления. При создании мер используется язык формул Data Analysis Expressions (DAX), включающий более чем 200 функций, операторов и конструкций. Например, возможна настройка динамических мер, которые будут вычисляться в зависимости от выбранного значения в фильтре (см. рис. 1).

Таблица 1. Вспомогательная таблица «Фильтр»

Параметр	N
Абсолютное число	1
На 100 тысяч населения	2

Для визуализации, представленной на рисунке 3, была создана дополнительная таблица «Фильтр» со значениями для фильтра «Единицы вывода». Далее была создана мера «Что отображать в столбцах», которая в зависимости от выбранного значения выводит на график одну из двух мер: «Абсолютная заболеваемость» либо «На 100 тыс. эпид.».

Формула меры «Что отображать в столбцах»:

```
Что отображать в столбцах = SWITCH(
true(),
VALUES('Фильтр'[N])=1,'Заболеваемость'[Абсолютная заболеваемость],
VALUES('Фильтр'[N])=2,'Заболеваемость'[На 100 тыс. эпид.],
BLANK()
)
```

Формула меры «Абсолютная заболеваемость»:

```
Абсолютная заболеваемость = SUM('Заболеваемость'[Заболеваемость])
```

Аналогичные меры были созданы для эпидемиологических показателей: заболеваемость и распространённость.

Рис. 1. Динамические меры

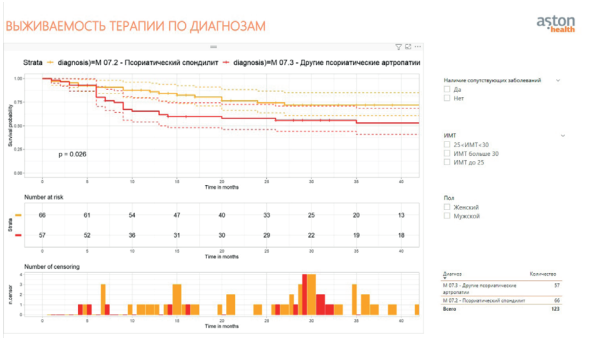

Рис. 2. Анализ выживаемости [6]. Сравнение выживаемости терапий в зависимости от диагноза. Визуализация на R

В Power BI предусмотрена возможность выбирать дополнительные визуализации из коллекции «Визуальные элементы Power BI». Помимо этого при настройке визуализаций есть возможность создавать визуальные элементы Python и R, что позволяет использовать различные библиотеки для статистической обработки данных и работы с графикой. А система взаимодействия между элементами отчёта предоставляет возможность быстро проверять большое количество статистических гипотез (рис. 2).

Для визуализации, представленной на рисунке 4, используется следующий скрипт, на основе [7].

```
library(survminer)
library(survival)
fit <- survfit(Surv(dataset$time, dataset$status) ~ as.factor(dataset$diagnosis), dataset)
ggsurvplot(

fit,                    # survfit object with calculated statistics.
fun = NULL,
dataset,                # данные, используемые для кривых выживаемости
risk.table = TRUE,      # показать таблицу рисков
pval = TRUE,            # показывать p-value на основе log-rank теста.
conf.int = TRUE,        # показывать доверительные интервалы для кривых выживаемости
font.legend = 16,
xlim = c(0,40),         # ограничения для оси X, не влияют на оценки выживаемости
palette = c("orange", "red", "green"),
xlab = "Time in months",  # подпись для оси X.
break.time.by = 5,      # интервалы для оси X.
ggtheme = theme_light(), # настройка графика и таблицы рисков с помощью темы
risk.table.y.text.col = T,  # цветные текстовые аннотации таблицы рисков
risk.table.height = 0.25,   # высота таблицы рисков
risk.table.y.text = FALSE, # показывать столбцы вместо имён в текстовых аннотациях
                        # в легенде таблицы рисков
ncensor.plot = TRUE,    # график цензурированных объектов в момент времени t
ncensor.plot.height = 0.25,
conf.int.style = "step",  # настройка стиля доверительных интервалов
surv.median.line = "hv", # горизонтальная и вертикальная линия для медианы
                        # выживаемости)
```

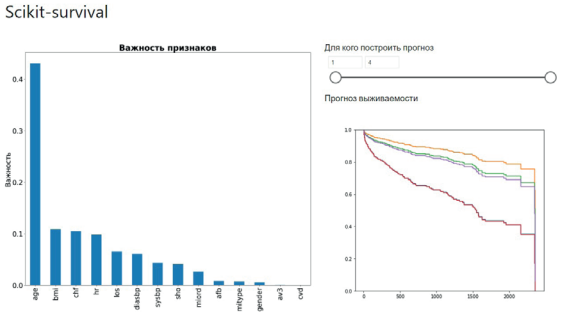

Рис. 3. Прогнозы выживаемости на основе входных параметров

Аналогично можно использовать библиотеки Python, в том числе для машинного прогнозирования. Например, метод «Случайный лес» для оценки выживаемости (Random survival forests), метод «Градиентный бустинг деревьев решений» (Gradient boosting trees) [8], которые представлены в библиотеке scikit-survival, автором которой является Sebastian Pölsterl [9, 10].

На примере открытых данных load_whas500 Вустерского исследования сердечного приступа, основной целью которого было описание факторов, связанных с тенденциями с течением времени в заболеваемости и выживаемости после госпитализации по поводу острого инфаркта миокарда [11], видно, что на основе созданной модели можно строить прогнозы выживаемости по входным параметрам (рис. 3).

Код для визуализации в Power BI важности признаков с помощью библиотеки scikit-survival:

```
import matplotlib.pyplot as plt
from sksurv.datasets import load_whas500
from sksurv.ensemble import GradientBoostingSurvivalAnalysis
from sksurv.ensemble import RandomSurvivalForest
import pandas as pd
X, y = load_whas500()
X = X.astype(float)
estimator = GradientBoostingSurvivalAnalysis(loss="coxph").fit(X, y)
plt.figure(figsize=(35, 15), dpi= 100, facecolor='w', edgecolor='k')
feature_imp = pd.Series(estimator.feature_importances_,
                        X.columns).sort_values(ascending=False)
fig, ax = plt.subplots(figsize=(16,14))
feature_imp.plot.bar(ax=ax)
ax.set_title("Важность признаков", fontsize=30, weight = 'bold')
ax.set_ylabel('Важность', fontsize=25)
plt.tick_params(axis='both', labelsize=25)
fig.tight_layout()
plt.show()
```

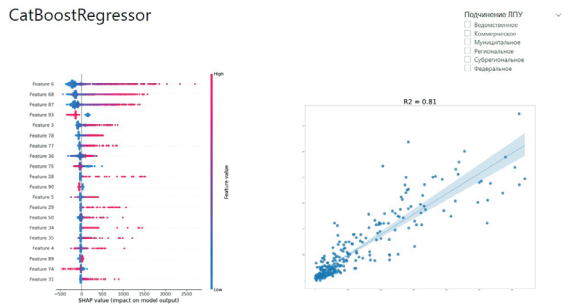

Рис. 4. Обработка данных с использованием библиотек Python: Shap и CatBoost

Так как для обработки скриптов используются установленные приложения на рабочем месте ис-

следователя, то можно использовать все библиотеки, установленные в них (рис. 4).

Обновление данных

Power BI позволяет настроить ежедневное обновление, обеспечивая доступ к локальным данным без необходимости ручного обновления. Пример отчёта, настроенного описанным выше способом, представлен на сайте: регистры эндокринопатий под эгидой ФГБУ «НМИЦ эндокринологии» МЗ РФ [12] (рис. 5).

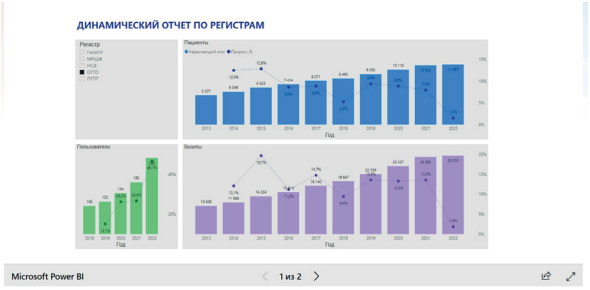

Рис. 5. Отчёт Power BI «Динамический отчёт по регистрам эндокринопатий» на сайте http://www.diaregistry.ru/

Заключение

Накопленный опыт в области применения средств бизнес-аналитики, в том числе Power BI, для обработки медицинских данных, полученных из различных источников, и их дальнейшей визуализации позволяет предложить его использование как дополнительного средства для обработки данных реальной клинической практики (RWD) и доказательств из реальной клинической практики (RWE).

ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ

Конфликт интересов. Автор работает в ЗАО «Астон Консалтинг».

Одобрение Комитетом по этике. Обзорная статья не требовала одобрения Комитетом по этике.

СВЕДЕНИЯ ОБ АВТОРЕ

Дмитриева Наталия Юрьевна — к.б.н., начальник отдела информационных систем ЗАО «Астон Консалтинг», Москва, Россия

e-mail: n.dmitrieva@aston-health.com https://orcid.org/0000-0003-1072-243X

Литература

1. Колбин А.С., Белоусов Д.Ю. Краткий отчёт о развитии доказательств, основанных на данных реальной клинической практики (RWD/RWE) в 2021 году: США, Россия и Евразийский экономический союз (ЕАЭС). *Реальная клиническая практика: данные и доказательства*. 2022;1(2):2–10. Доступно: https://doi.org/10.37489/2782-3784-myrwd-6.
2. Иванов А.В. Регистры как основа для сбора данных и построения доказательств. *Реальная клиническая практика: данные и доказательства*. 2021;1(1):10–15. Доступно: https://doi.org/10.37489/2782-3784-myrwd-3.
3. https://powerbi.microsoft.com/ [Internet] (дата обращения: 23.02.2022).
4. https://docs.microsoft.com/ru-ru/powerquery-m/ [Internet] (дата обращения: 23.02.2022).
5. Черников М.В., Новикова А.Н., Мазуров Н.Я. и др. Свидетельство о государственной регистрации программы ЭВМ № 2016615129, правообладатель АО «Астон Консалтинг» Доступно: https://new.fips.ru/registers-doc-view/fips_servlet?DB=EVM&DocNumber=2016615129&TypeFile=html.
6. Гланс С. Медико-биологическая статистика. — М.: Практика, 1999. — 459 с.
7. https://rpkgs.datanovia.com/survminer/.
8. Птушкин В.В., Мюллер М. Анализ эффективности лечения множественной миеломы на базе клинического опыта европейских стран. *Терапевтический архив*. 2021;93(4):404–414. doi: 10.26442/00403660.2021.04.200682
9. https://k-d-w.org/ [Internet] (дата обращения: 23.02.2022).
10. Pölsterl S. Scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *Journal of Machine Learning Research*. 2020;212(21):1–6.
11. Worcester Heart Attack Study data from Dr. Robert J. Goldberg of the Department of Cardiology at the University of Massachusetts Medical School.
12. Регистры эндокринопатий под эгидой ФГБУ «НМИЦ эндокринологии» МЗ РФ. Доступно: http://www.diaregistry.ru/ [Internet] (дата обращения: 23.02.2022).