PhUSE 2017

Paper IS05

Growing up with SAS: It Gets Real

Yevhen Los, Experis Clinical, Kharkiv, Ukraine
Dmytro Hasan, Experis Clinical, Kharkiv, Ukraine

ABSTRACT

In this paper we will explore one particular example of the real-world data and apply statistical analysis in order to uncover some of the mysteries it holds. Namely, we will take a look at the data about adverse events which occurred after the administration of the vaccines, collected by the Food and Drug Administration and the Centers for Disease Control.

We will see how even the most basic tools of data analysis in SAS® can provide valuable insights into the data, and how easily those "simple" tools can produce quite complex results. At the same time, we will employ more elaborate analyses in order to make some clear and easily interpretable conclusions.

We hope that our little example will inspire others to dig deeper into the data around us and search for the hidden gems and insights it can provide — with SAS it is not that difficult!

INTRODUCTION

It becomes increasingly recognised that accessing the benefits of a drug or therapy requires the knowledge on "real-world" outcomes. While Randomised Control Trials happen in idealized environment and measure efficacy in limited populations, real-life setting can differ significantly from this conditions, and it is very important to understand how the drug behaves in actual clinical practice. This paper describes the benefits of using the real-world data, as well as issues with its appropriate collection and reliability. The authors illustrated this information by analysing one particular example of such data, from The Vaccine Adverse Event Reporting System (VAERS), which is a free access database, created by the Food and Drug Administration (FDA) and Centers for Disease Control and Prevention (CDC). This data contains records of vaccination-related adverse events, and the authors analysed the population of children 12 through 21 month old. They started with downloading raw data from the database, and went all the way to the statistical analysis using some pretty basic tools of SAS software.

WHAT IS REAL WORLD DATA

Real World or Real-Life Data (RWD) in healthcare can be defined as any data gathered under real world practice circumstances, that are not collected in conventional Randomised Clinical Trials (RCTs). While RCTs remain the gold standard of accessing the benefits of a drug, it becomes increasingly recognised that measuring their impact in practical, real-life setting is also extremely important. RCTs by design are carried out using only selected population under idealised conditions, and thus only allow to measure the efficacy, but not the effectiveness of the drug. Real World Data allows to estimate effectiveness in typical practice conditions, and has other benefits:

- Allows to estimate the effects of a drug in a diverse study population that closely resembles the distribution of patients in real clinical practice;
- Allows to collect data when it is not possible to run an RCT (e.g. data on narcotic abuse);
- Allows to estimate long-term risk-benefit profile of a drug, including rare clinical benefits and harms;
- Allows to collect broader range of results than those usually collected in RCTs: Patient Reported Outcomes, Health-Related Quality of Life, etc.

RWD also can be used to obtain interim evidence which can be used to make some preliminary conclusions in the absence of RCT data, or at least to gain insights and formulate the hypotheses about the properties of the drug.

RWD can be collected from multiple sources, including:

1. Supplements to traditional RCTs. Such data can show treatment patterns for common events, e.g. doses of drugs used to treat one particular condition. The limitation of such evidence are same as those for RCTs: data is collected in same carefully selected population and clinical surrounding;
2. Large simple trials, also called practical or pragmatic clinical trials. They are by design larger than conventional RCTs, and have a benefit of randomisation, which minimises bias in estimation of treatment results. Thus they are more likely to capture significant outcomes in key areas of interest, applicable to more diverse population. The flip side of the coin is that they are relatively expensive and demand more control of the quality of data collected;
3. Health surveys and reviews. They usually collect the data on the representative individuals in the target population, and not just those who opted to participate in an RCT or chose a particular health plan. Their drawback is that they are usually extremely prone to subjectivity and recall bias;

4. Registries — prospective observational cohort studies on patients who receive a particular treatment for a particular condition, usually at a particular centre. The population of registries is usually more diverse than that of phase III of RCTs. However, they have all the limitations of the observational studies, such as lack of randomisation and questionable data integrity;
5. Electronic health records and medical charts. They may collect very detailed information at the personal level, both general and disease-specific. However, surveys and reviews data usually demand sophisticated analysis tools and may be biased by underreporting;
6. Administrative databases. They can provide huge volumes of data in short time. However, good quality and representative databases are still not as common as we would like, in addition to them having all the disadvantages associated with the retrospective analysis.

It is clear that if used correctly, RW data may provide clear advantage for understanding the outcome of treatment on all stages of the drug development, if collected and used correctly, and it is impossible to ignore the benefits of different forms of data collection in different situations.

Presently there is a wide selection of software that can be used for data manipulation and analysis, such as R, Python, SAS, etc. Virtually anyone who has at least some experience with any of the aforementioned tools can analyze the data they are interested in, and maybe even draw some conclusions from it. As the time goes, more and more real world data becomes available, and all of it can be processed and analyzed.

Since the authors predominantly use SAS in their work, we decided to stay faithful to it over the course of this article too.

VACCINATION-RELATED ADVERSE EVENTS

All of us have interacted with vaccination at some point of our life, we may remember how they were administered to us or to our children. Roughly speaking, vaccination is the administration of antigens in order to provoke the immune system into developing the immunity to the disease, which may prevent the disease altogether or at least mitigate the symptoms.

In this article we will work with the data from The Vaccine Adverse Event Reporting System (VAERS), which is a free access database, created by the Food and Drug Administration (FDA) and Centers for Disease Control and Prevention (CDC). You can download the datasets for the analysis in .csv format using this link: https://vaers.hhs.gov/data/index. For more details on how the database is organized and about the variables in the datasets please see VAERS Data Use Guide at https://vaers.hhs.gov/data/datasets.html.

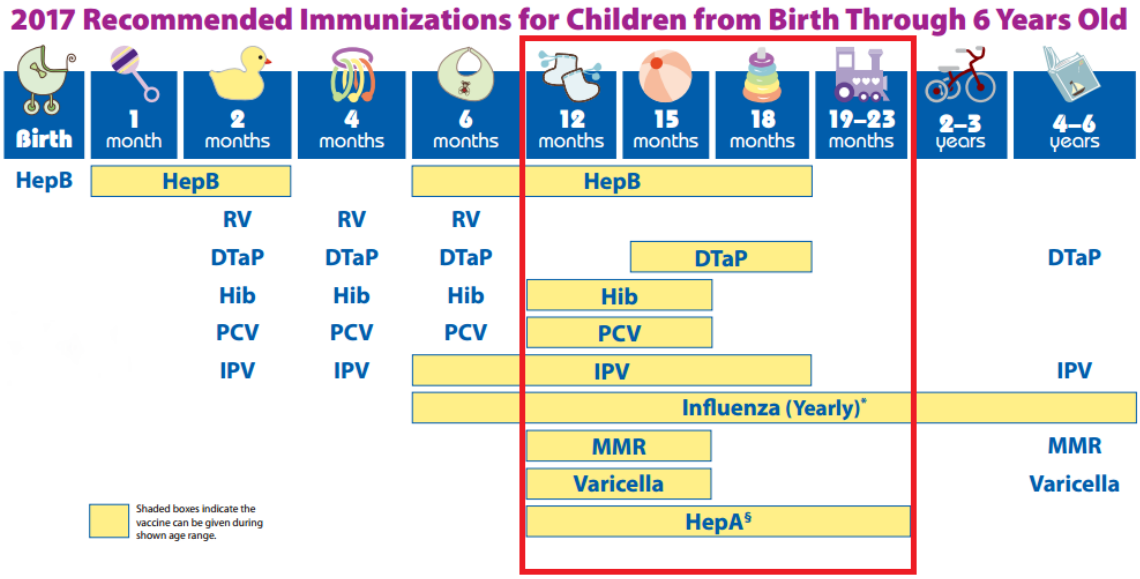
The authors of this article were interested in whether there is a measurably strong connection between the number of vaccinations and number of the adverse events that occur after them. Of course, there are procedures that are recommended by the doctors before vaccinations, and naturally, the manufacturers run trials and strive to make their products as safe as possible. In this article we just share the results of our research and let everyone make their own conclusions.

DATA STRUCTURE AND MANIPULATION

Let us take a look at the raw data and see what we have to offer. We downloaded the VAERS data which were reported from 2014.01.01 till 2017.01.13. Specifically, we examined the list of vaccines included in Immunization Schedules for Infants and Children (for more details see https://www.cdc.gov/vaccines/schedules/easy-to-read/child.html). A macro program was used to access the database and create the SAS datasets for our analysis. It reads the raw .csv datasets via IMPORT procedure. Since raw .csv files have column names in the first row, it suits our needs perfectly and saves us the trouble of using the INPUT statement. However, for more complicated or less organised data that might be unavoidable.

There were two datasets we are interested in: 20XXVAERSDATA.csv contains AE-related info, and 20XXVAERSVAX.csv describes the vaccines associated with the adverse events. Every observation from 20XXVAERSDATA.csv corresponds to a record from 20XXVAERSSYMPTOMS.csv with one or several AEs. Further for simplicity we will refer to one record from 20XXVAERSDATA.csv dataset as to an “AE” or “VAE”, since it describes a particular condition the patient manifested after receiving a dose of vaccine.

In this paper we were focusing on the specific vaccines, recommended for administration for children 12 through 21 month old (hence our “growing up” part of the title). We concentrated on this period of patient’s life as it’s probably the most vaccination-busy time for a person ever. Just take a look:



As such, in the subsequent datasteps we selected only relevant records from vaccines dataset (we used Immunization Schedules for Infants and Children for references to which vaccines are administered against which diseases). We then created variable DISEASE by selecting the names of the diseases from Immunization Schedules and corresponding them to our vaccines using page 11 from VAERS Data Use Guide.

It's worth mentioning that we checked that Immunization Schedules for Infants and Children did not change during our time interval, so all children were in more or less equal conditions.

After the processing of 20XXVAERSDATA.csv and 20XXVAERSVAX.csv we obtain data2014_2017.sas7bdat и vaccine2014_2017.sas7bdat which we use for the further analysis.

It is always a good idea to check the unfamiliar data for possible issues. In our case we had to filter out some duplicate records from our final datasets.

SERIOUS VAES

All VAEs matter, but in our analysis we decided to pay special attention to a class of events that we considered to be more notable. After all, is not out of the ordinary to have a fever or injection-site reaction after the vaccination. We considered an event to be more serious if one of the following was true:

- Patient died;
- AE was classified as life-threatening;
- Patient visited Emergency Room or was visited by a doctor;
- Patient was hospitalised.

In order to flag such events we created variable SEVAE in data2014_2017.sas7bdat. While the raw data does not make a clear distinction between “serious” and other VAEs, reports about such events do receive more careful scrutiny by the VAERS staff.

VAES AND NUMBER OF VACCINATIONS

We selected AEs which occurred after patients were administered several vaccines at once, and compared that data to the AEs which happened after patients received one vaccine exactly.

In order to classify our AEs we used the following SAS code:

```
proc sql FEEDBACK noprint;
    create table VAERS_IDS as
        select data.* , case when .<vac.N_TAKEN_V <=1 then 'One'
                        else 'Multiple' end as TAKEN,
        input(calculated TAKEN, ? obs_taken.) as N_TAKEN
    from lib.vaccine2014_2017 as data
        natural left join
        (select VAERS_ID, YEAR, count( distinct DISEASE) as N_TAKEN_V
            from lib.vaccine2014_2017 group by VAERS_ID, YEAR
        ) vac ;
quit;
```

At first we simply compared the numbers of events via PROC FREQ, and at the first glance it seemed that there were more VAEs associated with several taken vaccines. We decided to check if those numbers can tell us something decisive about our data. In order to do that we called PROC FREQ with ODDSRATIO and RELRISK options. ODDSRATIO shows the measure of association between our two groups of AEs in the output of PROC FREQ, and RELRISK statement which displays relative risk measures and confidence intervals for them, and p-values associated with the hypothesis that relative risk is significantly different from 1. We obtained the following result:

Association between serious VAEs and number of taken vaccinations.
Population: Infants aged 12-21 months.

Obs	Vaccines	Chi-Square	Cramer's V[2]	Odds Ratio[3] (95% CI)	p-value[1]
1	Chickenpox	61.70	0.24	4.260 (2.905 , 6.248)	<.0001
2	Chickenpox,Measles,Mumps,Rubella	0.88	0.06	1.324 (0.735 , 2.385)	0.3500
3	Diphtheria,Haemophilus b,Pertussis,Polio,Tetanus	0.44	0.09	1.530 (0.434 , 5.401)	0.5100
4	Diphtheria,Hepatitis B,Pertussis,Polio,Tetanus	2.43	0.16	2.799 (0.737 , 10.626)	0.1200
5	Diphtheria,Pertussis,Polio,Tetanus	12.39	0.26	4.370 (1.835 , 10.407)	<.0001
6	Diphtheria,Pertussis,Tetanus	10.94	0.12	1.909 (1.296 , 2.812)	<.0001
7	Haemophilus b	57.97	0.27	13.428 (5.793 , 31.124)	<.0001
8	Hepatitis A	6.45	0.09	1.693 (1.124 , 2.550)	0.0100
9	Hepatitis B	4.44	0.24	4.833 (1.001 , 23.344)	0.0400
10	Influenza	22.61	0.21	2.871 (1.840 , 4.482)	<.0001
11	Measles,Mumps,Rubella	15.53	0.12	1.873 (1.367 , 2.567)	<.0001
12	Pneumococcal	11.41	0.12	2.511 (1.449 , 4.350)	<.0001

[1]Corresponding p-value for Chi-Square statistic.
[2]the strenght measure of the assosiations that the Chi-Square test detected.
[3]the odds of SEVAE vaccination when it was received multiple vaccines to one vaccine.

Our data shows statistically significant difference in the number of VAEs recorded after taking several vaccines. For some vaccines there are several times greater numbers of VAEs appearing if they were administered more than once. While our data does not allow us to establish a cause-effect relationship, this result agrees with a common logic and can be considered as a food for thought.

VAES AND COMBINATION VACCINES

Sometimes the hypotheses are not as easy to formulate. Anyone can guess that vaccines are not administered several times unless something goes wrong, but is there something more interesting in the data? We decided to check the association between the number of AEs occurred after the administration of one vaccine against several diseases, and AEs that occurred after administration of several vaccines, while there was a corresponding combination vaccine.

Just as in the previous example, we simply asked PROC FREQ to compute a chi-square test. You can take a look at the result below:

Statistics for Table of Dise1 by TAKEN

Statistic	DF	Value	Prob
Chi-Square	8	458.8823	<.0001
Likelihood Ratio Chi-Square	8	555.4226	<.0001
Mantel-Haenszel Chi-Square	1	344.7632	<.0001
Phi Coefficient		0.7700	
Contingency Coefficient		0.6101	
Cramer's V		0.7700	
WARNING: 22% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Sample Size = 774

As you can see, we have a warning here. In this case SAS cannot reliably use the chi-square test to compute the chi-statistic and other some statistics, since some levels of the data have less than 5 observations. We cleaned our data to only include representative groups of covariates, and the warning disappeared:

Table of Dise1 by TAKEN

Dise1	TAKEN		
	Multiple	One	Total
Chickenpox,Measles,Mumps,Rubella	161	69	230
Diphtheria,Haemophilus b,Pertussis,Polio,Tetanus	6	13	19
Diphtheria,Haemophilus b,Pertussis,Tetanus	71	7	78
Diphtheria,Pertussis,Polio,Tetanus	9	49	58
Total	247	138	385

Statistics for Table of Dise1 by TAKEN

Statistic	DF	Value	Prob
Chi-Square	3	96.3404	<.0001
Likelihood Ratio Chi-Square	3	100.5748	<.0001
Mantel-Haenszel Chi-Square	1	19.3297	<.0001
Phi Coefficient		0.5002	
Contingency Coefficient		0.4474	
Cramer's V		0.5002	

Sample Size = 385

This output shows that statistically significant association exists between the number of VAEs and a type of vaccine(s) administered to the children. Again, while we cannot postulate that our data or methodology was clean and unbiased, this conclusion still may serve as a starting point or as a result to check in more rigorous examination. The reader is encouraged to add or build on this research if they have any ideas to share. Source code, both raw .csv and processed datasets for the analysis can be downloaded from a publicly accessed Github project: https://github.com/dimagmehanic/Growing-up-with-SAS-it-gets-real.

CONCLUSION

Real-world data is increasingly recognized as an essential tool for evaluating the benefits of a drug. There are various types and applications for RWD, and different types of RWD may hold different values depending on the circumstances. While RCTs remain the gold standard for demonstrating the clinical efficacy of the drug in the restricted setting, RDW can greatly contribute to the evidence of the drug benefits.

In this work there could have been much more analysis which we either didn't think of or didn't include in this text. If you are interested in some more numbers, you might want to take a look at the BONUS folder on the Github. We also may have missed something in terms of data issues, which is a common problem when dealing with the unfamiliar data. Anyhow, the reader is encouraged to try the repeat or build on our analysis further, and if they can make use of our programs on Github, then all the better.

This was a very simple and straightforward example of real world data analysis. We hope that this work will demonstrate that real world data is not just a mysterious concept from the future, but a relatively easily accessible thing, often reachable with just a few clicks on the Internet, and that in future RWD will be even more widespread. Hopefully a time will come when everybody skilled enough will have a possibility to check or estimate the effectiveness of drug or therapy with their own analysis rather than blindly believe things they see in mass-media.

REFERENCES

1. The Vaccine Adverse Event Reporting System. URL: https://vaers.hhs.gov/data/index
2. VAERS Data Use Guide. URL: https://vaers.hhs.gov/data/datasets.html
3. Immunization Schedules for Infants and Children. URL: https://www.cdc.gov/vaccines/schedules/easy-to-read/child.html
4. Dmytro Hasan. My bag of SAS® lifehacks. URL: https://www.pharmasug.org/proceedings/2017/QT/PharmaSUG-2017-QT05.pdf
5. Garrison LP Jr1, Neumann PJ, Erickson P, Marshall D, Mullins CD. Using real-world data for coverage and payment decisions: the ISPOR Real-World Data Task Force report. URL: https://www.ncbi.nlm.nih.gov/pubmed/17888097
6. Lieven Annemans, Michael Aristides, Maria Kubin. Real-Life Data: A Growing Need. URL: https://www.ispor.org/News/articles/Oct07/RLD.asp

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Yevhen Los
Experis Clinical
Haharina Ave 43/2
61000, Kharkiv, Ukraine
+1 407 512 1006 ext. 2417
yevhen.los@intego-group.com

Dmytro Hasan
Experis Clinical
Haharina Ave 43/2
61000, Kharkiv, Ukraine
+1 407 512 1006 ext. 2432
dmytro.hasan@intego-group.com, dimagmehanic@gmail.com