PharmaSUG China 2019 - Paper MA-023

Technical hot spots for career development - An overview of hot topics in PharmaSUG and PhUSE conferences

Chunpeng Zhao, Beigene;
Wenfang Li, Boehringer Ingelheim (China) Investment Co., Ltd;

ABSTRACT

As a technique and experience driven population, improve Job skills are always important for programmers to better support the fast growth of pharmaceutical industry. However, even most intelligent people may still face the confusion about what skills to start with and which specific areas to develop towards. Presenters in the conferences, whom are experts in each specific areas, usually share the most fancy techniques and realistic experiences overcoming challenges in daily work. Through multiple analysis and techniques, this presentation will show the dynamics of PharmaSUG and PhUSE conferences. Hope this could somehow shade a light on the way. All presentation related webpages were obtained from PharmaSUG and PhUSE websites taking the benefits of data crawling techniques such as web crawler in R. From where we extracted interested information i.e. title, authorship, company etc. for all presentations between 2010 and 2019 using Perl Regular Expression Functions in SAS. It is assumed that presenters would have the most representative key works in presentation titles. Such that we performed further analysis, mainly frequency analysis on key words grabbed from presentation titles. Through the study, we are trying to dig out the story between programmer and job skill development in Pharma related Areas. Besides, we are seeking the answers to the questions: what knowledge are helpful as a statistical programmer, and which emerging techniques are becoming hotter and hotter in Pharmaceutical areas. We would like to outline a technique knowledge sketch displayed in conferences. We hope that this could somehow be helpful in career development.

INTRODUCTION

PharmaSUG is a software users group of life science and health research professionals focused on the application of technological solutions in data analytics and regulatory support. PhUSE is an independent, not-for-profit, healthcare-focused biometrics society. Each year they will hold professional conferences, which attracts professionals all over the world with amount of valuable topics. Presenters in the conferences, whom are experts in each specific areas, usually share the most fancy techniques and realistic experiences overcoming challenges in daily work. As a technique and experience driven population, improve Job skills are always important for programmers to better support the fast growth of pharmaceutical industry. However, even most intelligent people may still face the confusion about what skills to start with and which specific areas to develop towards.

We extracted all presentation and paper titles between year 2010 and 2019 from PharmaSUG and PhUSE websites, and analyzed paper counts by year and region. We then explored distribution of duplicate paper titles across conferences, followed by frequency analysis on keywords in paper titles. We are in an era of change, so we also compared frequency changes of keywords between 2014-2016 and 2017-2019. Through multiple analysis and techniques, this presentation will show the dynamics of PharmaSUG and PhUSE conferences. Hope this could somehow shade a light on the way.

EXTRACTED INFORMATION FROM WEBSITES

We used web crawler in R to download all webpages from PharmaSUG and PhUSE websites. From selected PharmaSUG conference proceeding webpages and PhUSE Achieve webpages, we extracted paper codes, titles, authors and company information. Below are the R codes used to crawl and download all webpages from PharmaSUG and PhUSE websites. All downloaded webpages were saved under R installation sub-folders. Note: the following R code does not work under VPN.

```
install.packages("Rcrawler")
```

```
library(Rcrawler)
Rcrawler(Website = "http://www.pharmasug.org", no_cores = 4, no_conn = 4)
Rcrawler(Website = "https://www.PhUSE.eu/", no_cores = 4, no_conn = 4)
```

The webpages could be recognized as text files. Thus, we could use the following sas code to read and write them into sas dataset. '&I' in loop represents number of a webpage.

```
%let workfold=U:\PharmaSUG;
   filename html&I "&workfold/&I .html";
   data html&i;
     length STRING $2000;
     infile html&I length=l lrecl=2000 end=eof;
     input STRING $varying2000.l;
 run;
```

Structured data in webpages, such as paper codes, titles, authors and company information were extracted using data manipulation techniques in SAS. Perl Regular Express Functions are very helpful on this task. Although most data could be extracted successfully through SAS code, some manual work on data cleaning were still needed due to existence of unexpected data structure and special symbols in webpages. Finally, we got an excel file with all information stored as below.

Meeting	Location	Year	Section	Paper Code	Paper Title	Author	Company
PharmaSUG	US	2019	Advanced Programming	AP-001	Get Smart! Eliminate Kaos and Stay in Control - Creating a Complex Directory Structure with the DLCREATEDIR Statement	Louise Hadden	Abt Associates Inc.
PharmaSUG	US	2019	Advanced Programming	AP-018	Ensuring Programming Integrity with Python: Dynamic Code Plagiarism Detection	Michael Stackhouse	Covance
PharmaSUG	US	2019	Advanced Programming	AP-038	One Macro to Produce Descriptive Statistic Summary Tables with P-Values	Rajaram Venkatesan	Cognizant Technology Solution

Table 1. Sample data

DESCRIPTIVE STATISTICS ON PHARMASUG AND PHUSE

3932 papers and presentations in total were downloaded, of which 1862 (47.22%) were from PhUSE and 2070 (52.67%) were from PharmaSUG. In 2018, PhUSE started to hold conference in US. In PhUSE downloads, 1520 (38.64%) were from EU conferences (year 2006 to 2018), and 342 (8.69%) were from US conferences (year 2018 to 2019). In PharmaSUG downloads, 287 (7.30%) were from China conferences (year 2014 to 2018), and 1783 (45.37%) were from US conferences (year 2010 to 2019) (Figure 1). Well, the overall amount of paper from each region are gradually increasing year by year (Figure 2). In year 2018, among all downloads, US contributes the largest amount of topics (Figure 3).

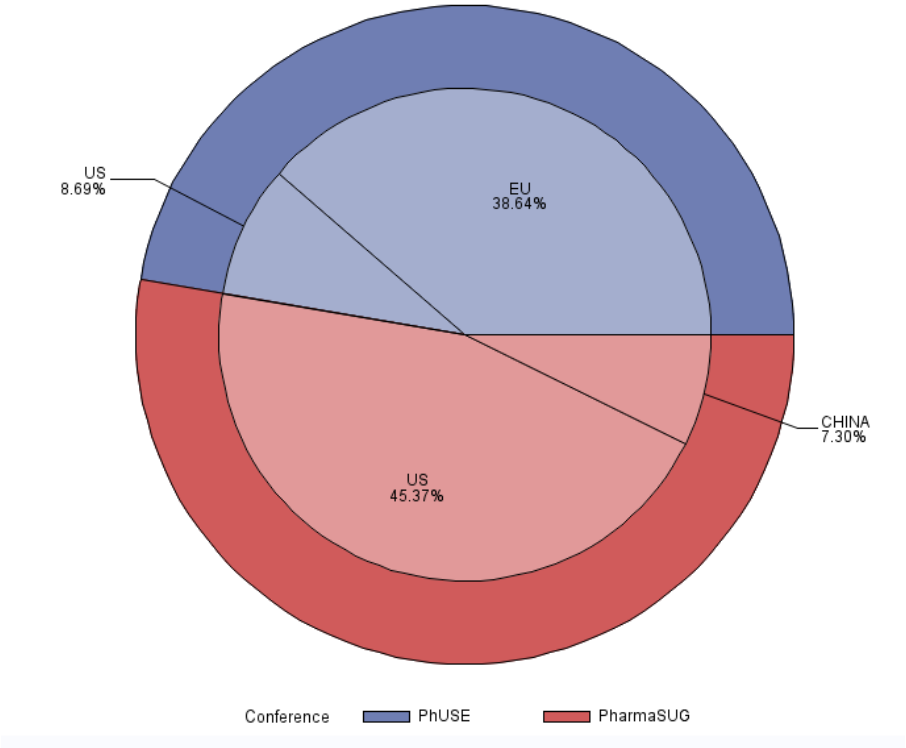

Figure 1. Conference Papers by Region

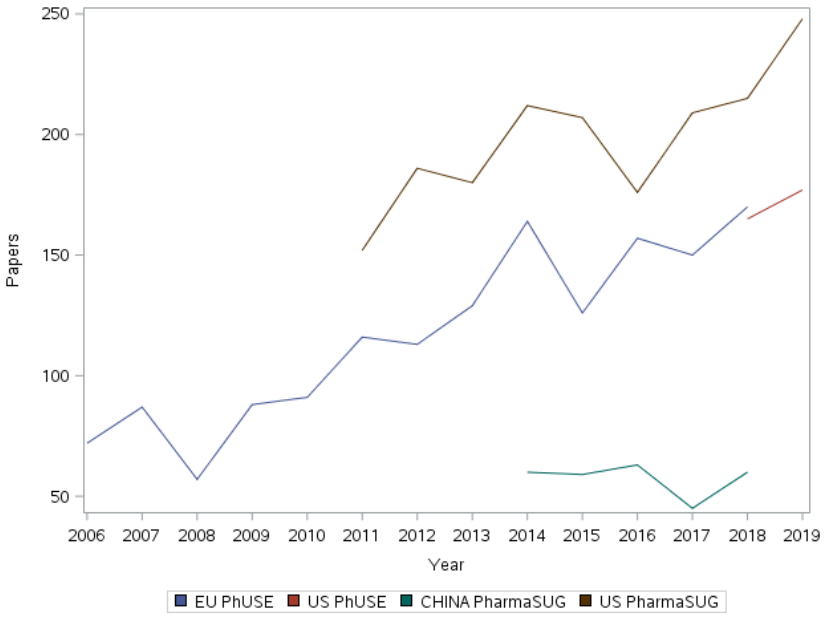

Figure 2. Conference Papers by Year

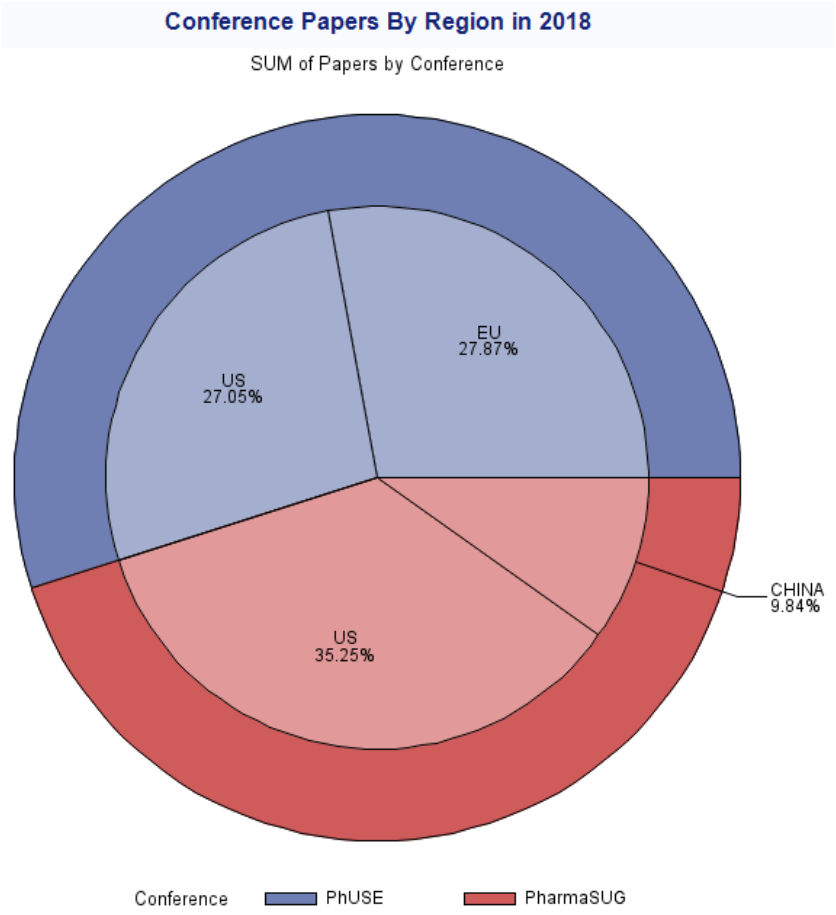

Figure 3. Conference Papers by Region in 2018

Among all presentations, most presentations have unique titles. Only 92 titles occur at least twice, which is in total 191 presentations (Figure 4). PharmaSUG China have 26 duplicate titles with PharmaSUG US, and no duplicate with PhUSE US/EU (Table 1). Some of these were illustrated from different aspects. Well, some were because same presentation was presented in multiple conferences. Well, the percentage of “duplicate” is less than 5% and even among those duplicate presentations, there were always new contents and thoughts. From some point of view, this may also indicates this group is very active in coming up new ideas, fighting with challenges and creating solutions.

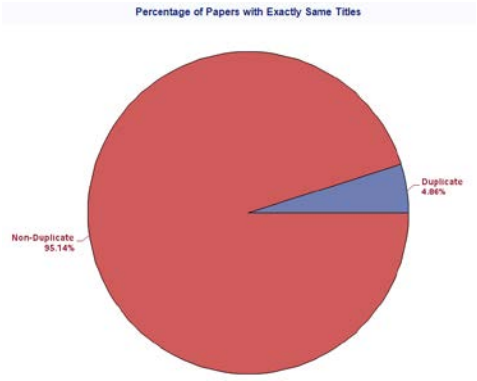

Figure 4. Percentage of Papers with Exactly Same Titles

	China PharmaSUG	US PharmaSUG	EU PHUSE	US PHUSE
China PharmaSUG	1	26	0	0
US PharmsSUG		17	8	15
EU PHUSE			16	6
US PHUSE				3

Table 1. Number of Same Paper Titles Shared between Conferences.

ANALYSIS TO KEY WORDS IN PAPER TITLES

Overall, there are 3834 unique presentation titles. To check the natural language features, we split all unique paper titles into individual words, extracted only meaningful verbs and nouns, and abandoned none-meaningful prepositions or conjunction words such as “the”, “a”, “an” and so on. Synonyms were grouped into one group. We then did frequency analysis on these words.

FREQUENCY ON ALL WORDS

From figure 5, we can see that, the most favorite noun in paper titles is “Data”, and its synonyms “Dataset(s)/Database” appeared in 932 paper titles. The most favorite verb is “Use”, and it appeared in 555 paper titles. SAS is still the most popular software used in those conferences for data analysis. If people would like to grab the most popular key words and assemble a sentence, it would be “Using SAS through Programming for Clinical Trial Data Visualization and Analysis”, which is also a very simplified job description for statistical programmer in pharmaceutical industry. This matches the very our common sense to the job in the past few years.

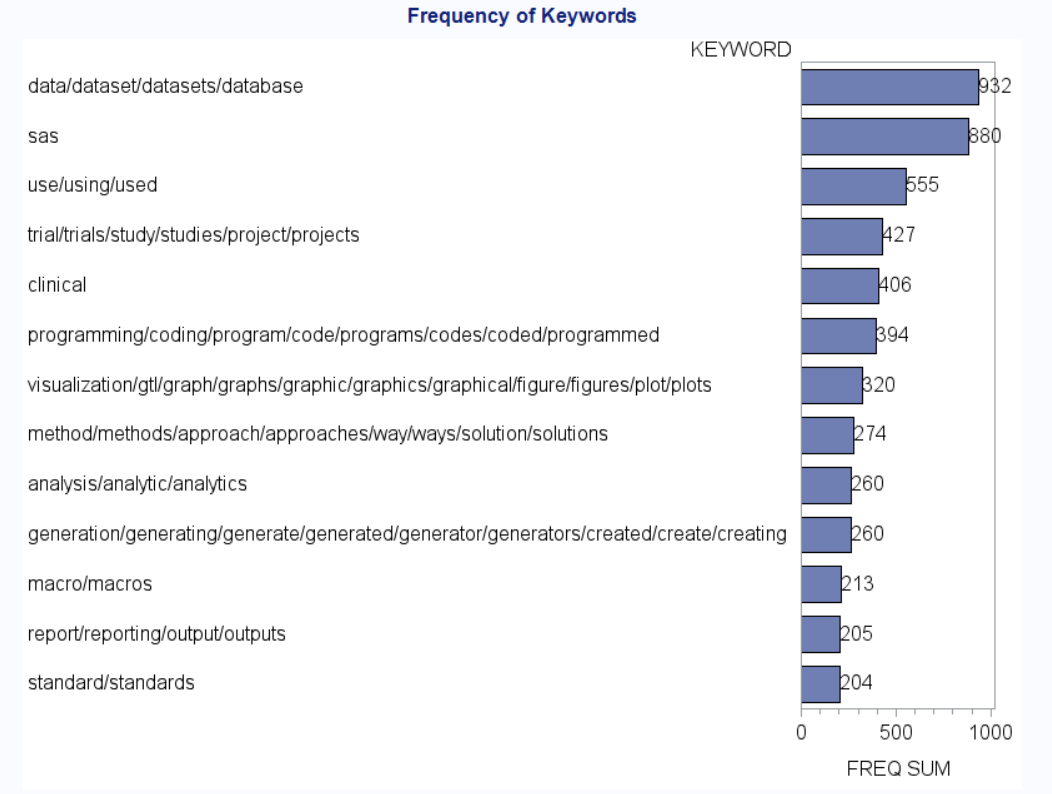

Figure 5. Frequency of All Keywords from Paper Titles

FREQUENCY ON SELECTED KEYWORDS

We also tried to select some keywords from the title keyword list to those illustrate frequently met challenges in our daily work. It is assumed that the more frequently the challenges will occur in daily work, the more likely it will appear in paper titles. Figure 6 listed the frequency of most frequent key words, from where we get the following:

1. "Data Visualization" were mentioned 320 times. This implies a strong business needs on data visualization and big challenges to all statistical programmers.
2. Statistical programmer tends to develop various macros and tools to facilitate work.
3. Automatic, dynamic work and interactive interface are among the top goals.
4. Follow CDISC standards such as SDTM, ADaM, Define files, and nicely meet submission related requirements are one of our big working challenges. It catches lots of eyes and takes lots of discussion.
5. Following SAS, excel stays at the 2nd position of software mentioned in the Keyword Frequency list. This indicated that, excel has been an essential assistant in work. There are 77 papers showing how Excel facilitates our work.
6. R is becoming more and more popular in the industry followed by python. Let's consider more about R implementations in the following section.

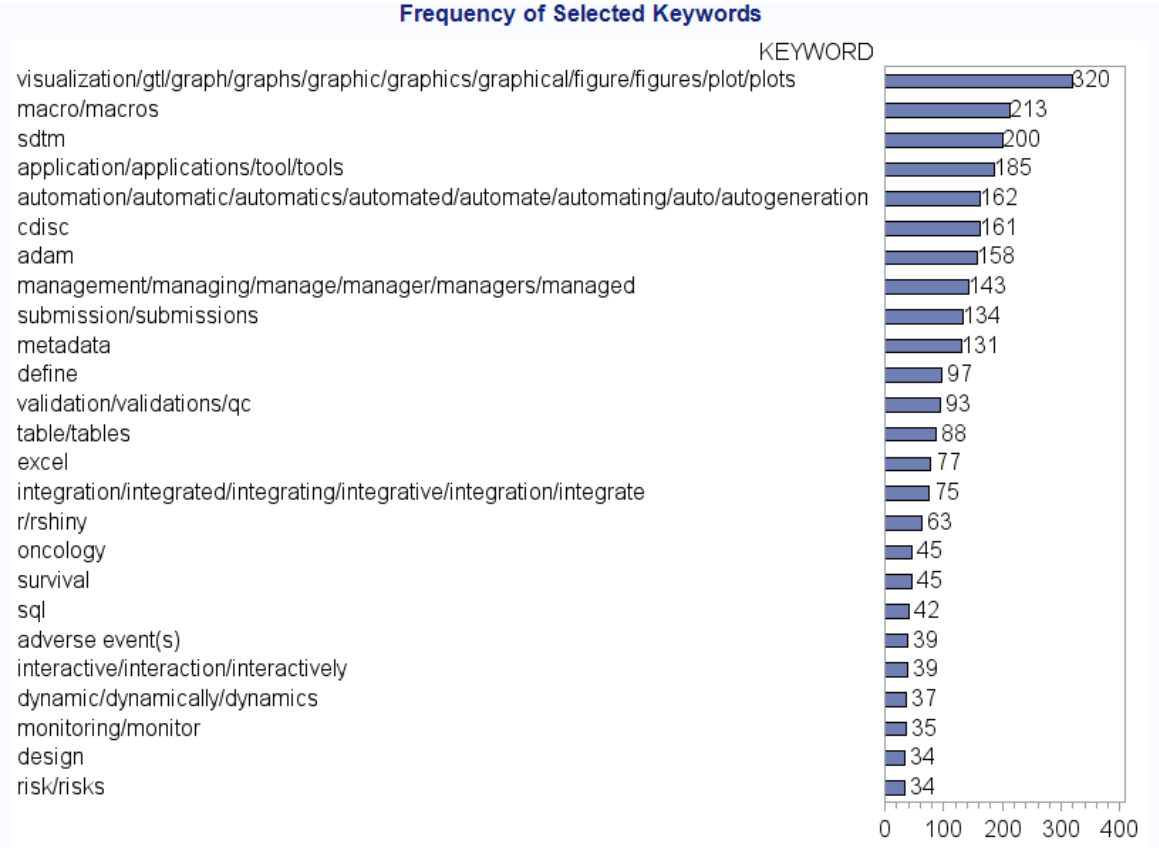

Figure 6. Frequency of Selected Keywords from Paper Titles

FREQUENCY CHANGES ON KEYWORDS BEFORE AND AFTER 2016

From 2014 to 2019, there are 2596 papers in total, including 1198 published during 3 years prior to 2017 (year 2014-2016) and 1398 during 3 years on and after 2017 (year 2017-2019). Clearly, there is a 200 increase in paper amount after 2017. Figure tells us:

1. R related papers increased by 30. The use of R is rapidly increasing in pharmaceutical industries.
2. Attention to programming language Python is growing in recent 3 years.
3. Submission and related CDISC standard topics, such as SDTM, ADaM and Define are becoming hotter and hotter.
4. Statistical programmers are investing more and more efforts to develop more automatic work and more interactive interface.
5. Data visualization is still our ongoing challenges.
6. New techniques have been introduced to industry, and people are actively exploring there implementation. In past 3 years, new fashionable keywords started to appear in paper titles, e.g., machine learning, artificial intelligence, blockchain and digital.
7. It seems we have reached a consensus on data transparency. Then we move on to data traceability. As one of the ADaM principles, data traceability is getting more attentions.
8. Data validation is still our ever-lasting challenges. Data qualities are highly depend on how we validate data.
9. How to analyze real world data, and how to monitor ongoing data are becoming a spotlight in the industry.
10. “SAS” and “Macro” seems to be disappearing from more and more paper titles, with R and Python rising. Maybe ten year later, R or Python programmers would become a substantial group in pharmaceutical industry, which could be comparable with SAS User Group.
11. Nonclinical data submission activities are increasing dramatically, according to “SEND” frequency change.

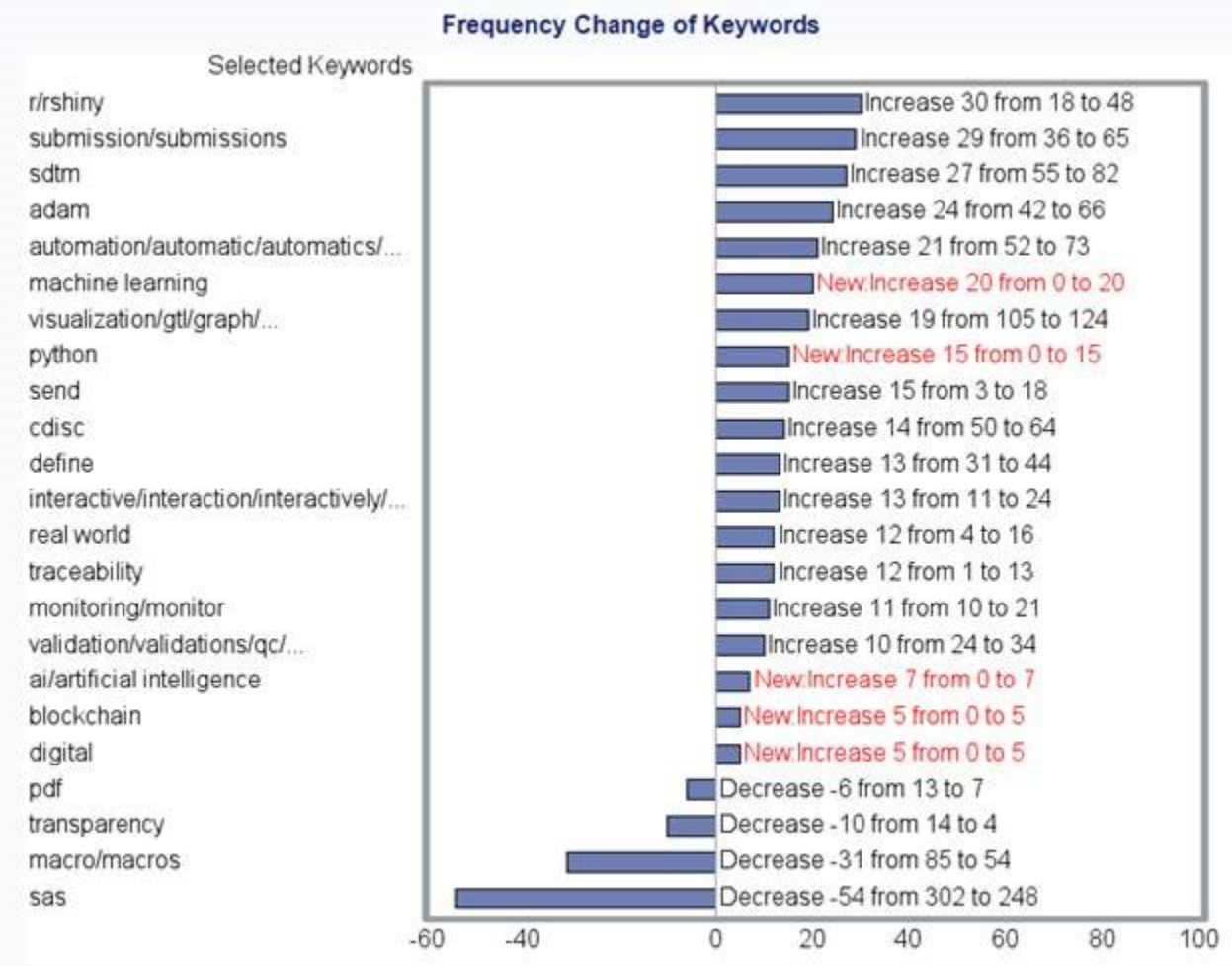

Figure 7. Frequency Changes of Selected Keywords from Paper from 2014-2016 to 2017-2019

CONCLUSION

Since the first US PhUSE held in 2018, PhUSE has attracted lots of papers in pharmaceutical areas all over the world. PhUSE conference is becoming a more and more important conference in pharmaceutical programming industry. China PharmaSUG is also taking more great paper and presentations with unique topics. R and Python is rising, distracting people substantial concentration on SAS. Besides R, programmers are also exploring more fancy technique implementations in pharmaceutical industry programming, e.g., machine learning, artificial intelligence and blockchain. There is no doubt that CDISC standards are still boiling topics.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. The excel file with all paper titles can be sent upon request. Contact the author at:

Name: Chunpeng Zhao
Enterprise: Beigene
Address: 20/F, Tower 3, Jing An Kerry Centre, 1515 Nanjing Road West
City, State ZIP: Shanghai, China
E-mail: chunpeng.zhao@beigene.com