PharmaSUG 2021-Paper LS-093

Why Data Scientists need leadership skills? Story of Cross-Value Chain Data Utilization Project

Yura Suzuki, Shionogi & Co., Ltd .;
Yuichi Koretaka, Shionogi & Co., Ltd .;
Ryo Kiguchi, Shionogi & Co., Ltd .;
Yoshitake Kitanishi, Shionogi & Co., Ltd.

ABSTRACT

Shionogi Data Science (DS) Office is a new organization established in April 2020. One of our mission is to contribute to planning and promotion, and to support from the aspects of statistics and data science across the value chain and to contribute to management decisions based on scientific evidence. Many members of the DS Office have lots of experience in statistical analysis, programming, and data management in clinical drug development. We are playing a new role as data scientists based on the skills that have cultivated up to now, while gaining further strengths and skills. We have more opportunities to interact with colleagues of each value chain, and we are promoting the solution of business issues based on various data. In addition to the ability to carry out data analysis, leadership skills are required such as the ability to identify and summarize their needs, the ability to draw out an analysis plan to solve the issues, the ability to carry it out in cooperation with the colleagues in the DS Office, the ability to show results in an easy-to-understand manner, and the ability to discuss next action within the team. In this paper, we will take up the "Cross-value chain data utilization project" as a concrete example and introduce the leadership skills required for mid-level data scientists.

INTRODUCTION

Shionogi Data Science (DS) Office is a new organization established in April 2020. Mission of DS Office is to contribute to drawing out and forwarding data-based strategy that makes full use of advanced analysis technology for enlightenment, prevention, diagnosis, treatment of diseases, and maintenance and promotion of health. Another mission is to contribute to planning and promotion, and to support from the aspects of statistics and data science across the value chain and to contribute to management decisions based on scientific evidence. The specific business contents of DS Office are as follows;

- Understanding internal and external data, and planning and promoting appropriate utilization
- Providing useful information for healthcare strategies and research plans by utilizing databases and setting hypotheses
- Development and utilization of methods for efficient data analyses and provision of information that can be used for decision making inside and outside the division
- Streamlining procedures and necessary computing environment for more efficient analysis activities
- In-house data literacy and data science education planning and promotion
- Consulting and supporting analyses from a data science perspective for issues that occur company-wide

Many members of the DS Office have lots of experience in statistical analysis, programming, and data management in clinical drug development. We are playing a new role as data scientists based on the skills that have cultivated up to now, while gaining further strengths and skills.
For example, Yura Suzuki, the first author of this paper, conducted data analyses of clinical trials while accumulating a career as a Statistician and SAS® Programmer in clinical trials of drug development for more than 10 years, while also being involved in the construction of an analysis programming platform and have acquired knowledge about the system. After moving to the DS Office, I have more opportunities to interact with colleagues of each value chain other than Clinical Development Department, and I'm promoting the solution of business issues based on various data, not limited to clinical trial data. In addition to the ability to carry out data analysis, leadership skills are required such as the ability to identify and summarize their needs, the ability to draw out an analysis plan to solve the issues, the ability to carry it out in cooperation with the colleagues in the DS Office, the ability to show results in an easy-to-understand manner, and the ability to discuss next action within the team (Figure 1). It is required for the leader to grasp each issue accurately, to imagine the future of solving those issues, and to gather

available resources (e.g., team members, data, environment) while accurately grasping each issue and imagining the future of solving those issues. It is required to promote the solution.

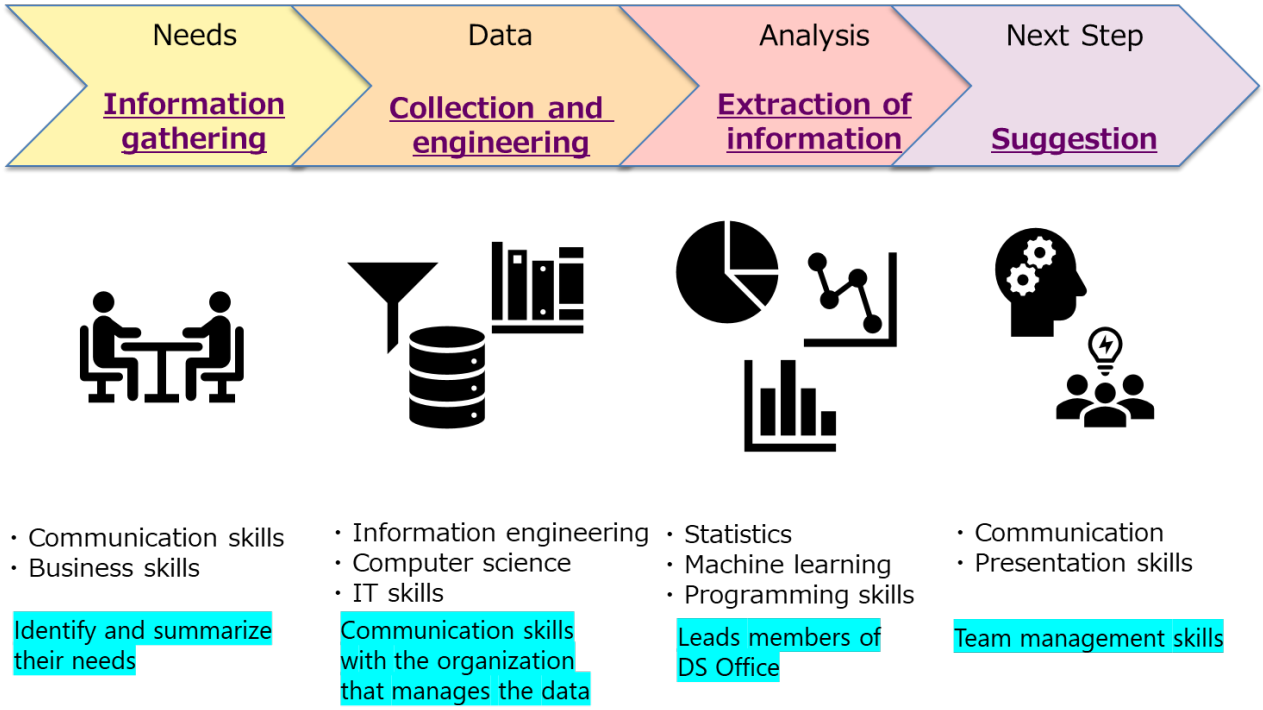

Figure 1. The role of data scientists in solving business problems. Light-blue highlighted skills are related to leadership skills.

Therefore, the skills and knowledges regarding data analysis and system environments gained in the past careers need to continue to step up and acquire the latest technology as a concrete means to solve problem, while the leadership skills have become even more important as the value chain of working together has expanded. In this paper, we will take up the "Cross-value chain data utilization project" as a concrete example and introduce the leadership skills required there.

EXAMPLE: CROSS-VALUE CHAIN DATA UTILIZATION

PROJECT OUTLINE

The cross-value chain data utilization project consists of members of each value chain of R&D and post-marketing (e.g., safety and drug information center), and members of the DS Office. The background to the launch of this project was the following various restrictions on the data, system and structure;

A) The data of each value chain is managed by each system validated under different GxP regulations, and so it is difficult to cross-reference those data and create new hypotheses (see Figure 2).

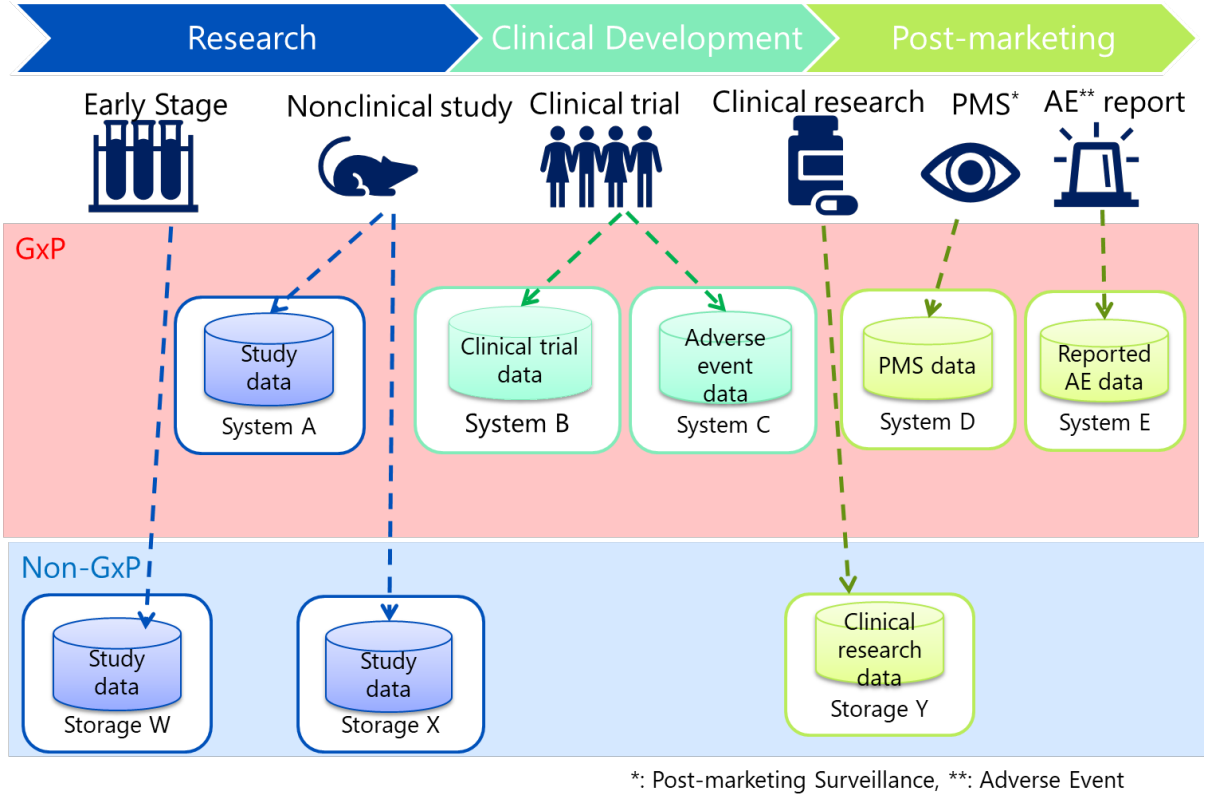

Figure 2. Data of compound/product scattered throughout Shionogi

B) The data managed by each organization can only be grasped by the members involved in the compound/product, and no one has a complete picture of what kind of data exists in which system in Shionogi. Even if he or she is a member involved in a compound, it will be difficult for the him/her in charge of the previous phase to refer to the data when the phase of the compound progresses. Therefore, it is difficult to utilize data across the value chain even for a single compound.

C) When multiple compounds/products are used for the same patient profile, it is desirable to utilize data across compounds and promote strategic planning on the disease across multiple compounds/products, but this hasn't been achieved due to problems (A) and (B).

D) Some of data analyses were outsourced, and in some cases there may be no opportunity to perform additional analysis in-house after the data is delivered from the contractor. So, there is a shortage of human resources who can perform data analysis.

In order to make maximum use of compound/product data in Shionogi and carry out strategic planning on the disease with multiple compounds/products, we launched a cross-value chain data utilization project.
In the project, three teams were established according to three diseases, and members of R&D and post-marketing (e.g., safety and drug information center) were assigned for each team. The members are familiar with the disease or the data held by each organization. In addition, four or five members of the DS Office were assigned to each team, and one of them was designated as the team leader to lead various members so that the team could formulate strategies. The challenges faced in leading the diverse members are shown in the subsequent section.
The team structure, process and timeline of this project is shown in Figure 3. Within six months, each of the three teams started from discussion of idea to discussion based on the results of data analysis. In addition, the progress and issues of each team were shared at the monthly Leaders Meeting to cooperate among the team leaders.

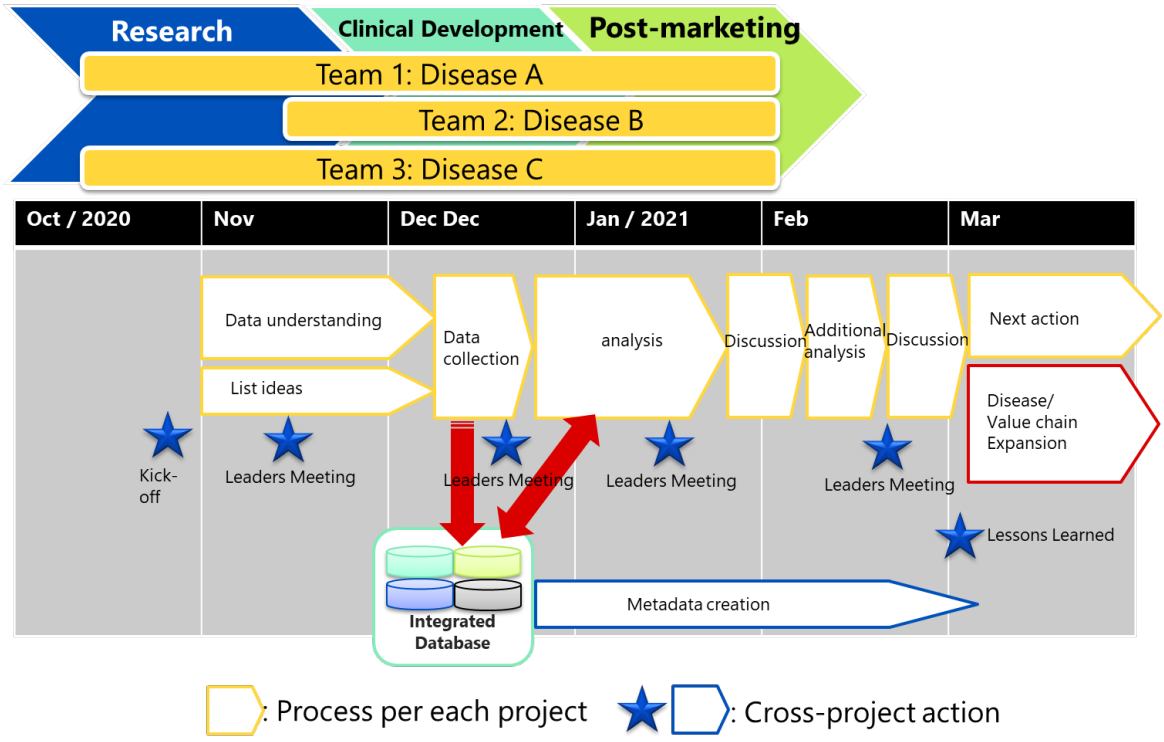

Figure 3. Team structure, process and timeline of this project

From here, we will explain how the leaders of each team demonstrated their leadership skills in each process shown in Figure 3.

DATA UNDERSTANDING / LIST IDEAS

In order to solve problem (B) in the previous section, we started with the step of grasping the data managed in each value chain. In handling data, the viewpoint of personal information protection is indispensable, and it is important to educate members about personal information protection in order to comply with it. When members use the Integrated Database environment described later, Shionogi conducts a test regarding personal information protection, and grants access to the Integrated Database environment after the member passes the test.
Some data was collected based on the plan, such as non-clinical/clinical trials and post-marketing surveillance (PMS), and was finalized when each study ended, and is basically structured data. On the other hand, some data is collected from various customers after marketing and is updated daily, and it is unstructured data. Even if the data was collected from the trial, some clinical trial data was collected in an earlier time, and there was a problem that the specifications of the analysis data set are not unified across trial since it was before the application of SDTM/ADaM in CDISC format.
In addition, since the SAS® format catalog created by the previous OS does not work in the current system environment. To overcome this problem, we used Shionogi Global SAS Server [1] to refer to the data in the environment of SAS® 9.2 and so we were able to confirm the contents of the data.

In the data understanding process, not only the leader but all the team members correctly understood the characteristics of each data, and then listed the ideas that we would like to verify using these data. After that, the team members prioritized the issues to be solved. At most, about 20 issues were raised per team, the process of prioritizing issues was necessary since the goal of this project was to make one or

more new hypotheses within 6 months. If there are various members of multiple value chains involved in the team in determining the issues, it may be difficult to prioritize them to the satisfaction of all, but the team leader organized the issues raised by the members and considered the following viewpoints. Therefore, we decided which task to start with the highest priority.

・ Data sufficiency

・ Relationship between tasks. For example, task A and task B can be tackled at the same time as one theme, task C should be solved before moving to task D, etc.

・ Versatility of the task (i.e., Is the task tackled within the team applicable to other diseases?)

・ Difficulty of the task (i.e., Will the analysis be completed in a period of about 1 to 2 months?)

Each team considered whether new hypotheses could be derived based on the issues that occur in the medical field at the stage of post-marketing. For example, one team decided to deep dive into patients that were discontinued after drug administration and the usage of concomitant drugs during administration with clinical trial data, PMS data, reported AE data, and Real World Data (RWD). Another team decided to deep dive into the relationships between Quality of Life (QoL) and other endpoints from past clinical trial data, and explore the association between meteorological condition and disease by using meteorological data and clinical trial data. The other team decided to perform text mining with internal text data to extract needs in the medical field.

The issues that were not selected in this process are still important, so as soon as the issues to be prioritized are completed, the team needs to consider the possibility of solving each issue again. The leader informed it each member.

DATA COLLECTION / ANALYSIS

Once the issues to be prioritized were decided, we moved to the process of aggregating the data required for data analysis. As shown in Figure 2, data is scattered in multiple systems owned by each organization, and it was necessary to consolidate the data to be used in one place before performing the analysis. To overcome this problem, we decided to copy data from each system to the Integrated Database System owned by the DS Office.

The Integrated Database System was constructed to aggregate and accumulate data collected in-house, open data and purchase data, and to utilize them. Data analysis environments such as SAS® Viya®, RStudio, and Python are in place, and it is possible to access databases and tables on the Integrated Database System from these environments to perform data analyses. From the viewpoint of personal information protection, data in the Integrated Database System is properly anonymized.

The team leader cooperated with the Admin of the Integrated Database, updated the progress of the project and the Database/Table that the team wanted to create in the Integrated Database frequently, and created an environment where team members could access the data. As mentioned in the "Data understanding / List ideas" section, in order for members to comply with the personal information protection, an operation to conduct a test of the personal information protection and grant system access rights to those who pass the test. Also, data access rights were strictly controlled according to the origin of the data. In addition, the team leader informed team members that the place where data was handled was limited to the Integrated Database environment.
Also, before starting data analysis, the team leader confirmed skills of data handling and data analysis for each member. The members assigned to the project had different backgrounds, such as those who were familiar with each disease and those who were familiar with how to collect each data, and each member had different data handling/analysis skills. As a result, the members of the DS Office who have the skills of programming by using SAS®, RStudio, and Python, performed the analysis mainly, and in some teams, the members other than the DS Office also performed data visualization and analysis by themselves. In addition, the other members visualized the data with data analytics tool such as Spotfire. As a result, we built an environment where all the team members got involved in the analysis process. In addition, the team leader summarized the analysis plan and assigned tasks of the members in the DS Office.

DISCUSSION

After performing the data analysis, the results were shared and discussed within the team. Before discussion, the members who performed the analysis explained the contents of the analysis to the members in the team in an easy-to-understand manner, and all members of the team interpreted the analysis results. To find new hypotheses based on the data, various members like researchers in the disease field, members in the post-marketing value chain which close to the medical field, reviewed and interpreted the analysis results.

As a result, one team examined patients who were discontinued after administration of the drug, found results that could be used in the future as safety information. That team also examined the actual usage of concomitant drugs during administration, and it was suggested that the usage of concomitant drugs may reduce the discontinuation of administration. Another team performed data analysis with past clinical trial data and it was suggested that there may relationships between QoL questionnaire and other several endpoints, providing insights that could provide a foothold in treatment strategies for the disease. The other team performed text mining using internal text data and gained a foothold for promptly providing information according to the interests of the medical field. In this way, we were able to find hypotheses based on data for the highest prioritized issues selected in each team.

Regarding the next action of each team, we need to verify to improve the accuracy of the obtained results by conducting analysis using other trial data, and also we and need to coordinate with related departments to build the flow to utilize the obtained results in the actual field.

LESSONS LEARNED

Members of various value chains in Shionogi participated in cross-value chain data utilization project. Since the data is scattered in multiple systems owned by various organizations, the leaders' first task was that we explained the purpose of data utilization to each organization and built this project so that necessary members can participate. At the same time, it was also important to make known that we would like to bring in ideas for data utilization from each value chain first. Some people were worried about being involved in the project without knowing what kind of idea is brought to verify by utilizing the data. In that case, we told them that we would like to realize the process of coming up with ideas firstly, and secondly collecting and analyzing necessary data, and developing arguments with various members from different backgrounds, since there is a limit to creating new hypotheses only in one value-chain.

After the project was launched, each team leader in the DS Office listened to the opinions of various members in the team, and we pulled the issues in each value chain together. Also, we gathered information of available data which managed by each organization comprehensively, and we investigated types and characteristics of each data and precisely. After that, we considered a plan to combine these various data to verify the issues. We played the role of managing each process according to the project timeline. Regarding the data analysis, we explained how to view the results to the members, summarized the discussions of the team, and conducted additional analysis necessary for further discussions. As a result, the team leader of the DS Office played the role of connecting each value chain, and it has become possible to carry out data-based verification of issues while collaborating with the members of each value chain.

On the other hand, the following issues have also become apparent;

Firstly, the shortage of human resources who can create databases and human resources who can analyze data was a serious problem. It is desirable that the database is being constructed so that analysis using the database can be carried out promptly once the issues to be prioritized are decided, but as mentioned in the "Data understanding / List ideas" section, there were a couple of problems with the data (e.g., how to handle SAS® format catalog created by the previous OS, how to create database from non-SDTM/non-ADaM). Due to these problems, it was not possible to build a database in a short period of time with limited resources, so this time we stored the data of each value chain in the area in the integrated database environment and analyzed it directly without creating a database. In the future, we

would like to increase the number of Database and Tables which are to be useful not only for this project but also for other projects.

Regarding human resources who can handle data analysis, not only members of the DS Office but also members other than the DS Office cooperated with the data analysis in some teams of this project, and it was able to promote the data analysis within the team efficiently. However, resources were very limited, and the members in charge of data analysis had to perform multiple data analyses in a short period, which was a heavy burden. Since it is necessary to proceed with the data analysis process in a timely manner, such as promptly conducting data analysis after deciding the issues to be prioritized and performing additional analysis that occurred in the process of discussion based on the analysis results. When there were multiple issues, the team leader created a couple of small subgroups within the team and devised a method to specialize in a specific issue and proceed with data analysis efficiently. However, in the future, the number of teams will be expanded if the target diseases are expanded. If so, the number of issues to be tackled in parallel at the same time will increase, and so it is necessary to develop human resources who can handle data analysis. Regarding the data to be handled, one idea is that the DS Office will develop an education as a case study through the case of this project so that members other than DS Office can correctly grasp the structure and characteristics of the data and can visualize the data by themselves. In addition, as we did this time, it is also necessary that the members who conducted the data analysis explained the appropriate view of the analysis result to other members before discussing within the team.

Secondly, due to restrictions on the data to be handled, there may be cases where discussions are biased toward the company's compounds, or when handling multiple types of data, it may be difficult to interpret when there is a discrepancy in the results between the data. In the future, it will be necessary to discuss whether we have to acquire new data or not, which data should be prioritized when there is a discrepancy in the results, and how we confirm the reproducibility using other available data.

CONCLUSION

As one example where leadership skills are required of data scientists, we introduced a case in which members of various value chains in our company participated and launched and promoted a cross-value chain data utilization project. In this project, the team leader selected from the DS Office played a central role in comprehensively and in detail grasping the internal data managed by each organization and the issues in each value chain, and using them for each issue. After understanding the types and characteristics of available data, we planned how to combine these various data, set a path for verification of issues, and performed data analyses. We managed the overall timeline of each process. Regarding the analysis performed, we explained how to refer the results to the members, summarized the discussions of the members, and conducted additional analysis which was necessary for further discussions. As a result, the team leader of the DS Office played the role of connecting each value chain, and it has become possible to carry out data-based verification of issues while collaborating with the members of each value chain.

The example of the cross-value chain data utilization project discussed this time is an example of how data scientists demonstrated their leadership skills and led the team. In the future, the cross-value chain data utilization project will solve the issues described in "Lessons learned" and will expand the scope of target diseases and value chains. This project is just an example why the data scientists need leadership skills – actually, there are many situations in which data scientists solve internal business issues. Needless to say, data analysis skills are indispensable for data scientists. It is necessary to continue to acquire and improve the skills to analyze various data, however, to carry out data analysis is just part of skills of data scientists. It is also necessary for data scientists to coordinate with departments that have business issues, to consider how to approach the issues using data, and to adjust the resources (i.e., team members, data, environment) required to solve the issues. Since it is increased to solve business issues by using various data, we believe that the leadership skills of data scientists will become even more important in the future.

REFERENCES

[1] Yura Suzuki et al. 2019. “SHIONOGI Global SAS System Renewal Project -How to improve Statistical Programming Platform-” Proceedings of the PharmaSUG 2018 Conference, Philadelphia, PA: PharmaSUG.

Available at https://www.pharmasug.org/proceedings/2019/SI/PharmaSUG-2019-SI-075.pdf.

ACKNOWLEDGMENTS

We would like to thank all the members who participated in each process of the cross-value chain data utilization project. The members have various specialties, and the team was given input from various perspectives, and they spent a lot of resources in this project for a limited period of time. We would also like to thank the heads of each organization for approving this activity.

CONTACT INFORMATION

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

Yura Suzuki
Shionogi & Co., Ltd.
+81-6-6485-5104
yura.suzuki@shionogi.co.jp

Yuichi Koretaka
Shionogi & Co., Ltd.
+81-6-6485-5104
yuichi.koretaka@shionogi.co.jp

Ryo Kiguchi
Shionogi & Co., Ltd.
+81-6-6485-5104
ryo.kiguchi@shionogi.co.jp

Yoshitake Kitanishi
Shionogi & Co., Ltd.
+81-6-6485-5104
yoshitake.kitanishi@shionogi.co.jp