PharmaSUG China 2025 - Paper AP-109

SemiSPRG: Semi-automated generation of Study Data Standardization Plan and Reviewer Guides in regulatory submission in R

Jinjing HE, Chasen CHEN, SANOFI

ABSTRACT

SemiSPRG is an R-based tool designed to semi-automate the generation of Study Data Standardization Plan (SDSP) and Review Guides (RGs) including Clinical Study Data Reviewer's guide (cSDRG) and Analysis Data Reviewer's guide (ADRG) in regulatory project submission. This semi-automated instrument optimizes the efficiency and precision of SDSP and RGs production by systematically extracting crucial information from various project components, including project outputs, project programs, SDTM or ADaM metadata files and pinnalce21 reports, inserting into the corresponding location in the SDSP and RGs' template, thus facilitating the creation of SDSP and RGs in Microsoft Word format. The R Shiny technology underpins the operational framework of SemiSPRG. Specifically, R Shiny provides an interactive interface for users, along with a platform for the upload of relevant project documents and the subsequent download of SDSP and RGs. The implementation of SemiSPRG is based on R version 4.4.1. R packages such as flextable, officer, xml2, etc. will be used to assist in document generation. By mitigating the traditionally labor-intensive nature of submission document preparation, SemiSPRG enhances workflow efficiency significantly, providing an invaluable resource for professionals requiring meticulous and expedited documentation processes in regulatory project submission.

INTRODUCTION

The compilation of submission packages for Health Authorities (HA) necessitates meticulous planning and the inclusion of standardized documentation. Among these, the Study Data Standardization Plan (SDSP), the Clinical Study Data Reviewer’s Guide (cSDRG), and the Analysis Data Reviewer’s Guide (ADRG) are critical components for outlining and contextualizing study data. The SDSP is expected to detail proposed Study Data Tabulation Model (SDTM) domain and variable utilization, supplemental domain implementation, and linkages to planned analyses. Concurrently, the cSDRG and ADRG must furnish pertinent information related to the SDTM and Analysis Data Model (ADaM) structures, respectively.

A substantial portion of the content within these documents can be directly derived from existing study assets, including SDTM/ADaM metadata, the datasets themselves, associated programming code, TFL outputs and reports from Pinnacle 21 validation tools. To enhance the efficiency of generating these essential documents, a software tool, SemiSPRG (Semi-automated SDSP, cSDRG, and ADRG Generation), has been developed. SemiSPRG is engineered to automatically extract and integrate information from these diverse study materials into designated sections of the SDSP, cSDRG, and ADRG. The tool incorporates an R Shiny interface, which facilitates user interaction for uploading the requisite source materials and subsequently downloading the partially auto-populated documents. This paper delineates the operational framework and methodology employed by SemiSPRG for the streamlined generation of the SDSP, cSDRG, and ADRG.

THE SEMISPRG TOOL: WORKFLOW AND IMPLEMENTATION

SemiSPRG is developed in R, leveraging the Shiny framework for its interactive user interface. The core principle of the tool is to populate predefined SDSP, cSDRG and ADRG templates (.docx) with study-specific information extracted from various input study files.

OVERALL WORKFLOW

The workflow of SemiSPRG can be summarized in the following steps (Figure 1):

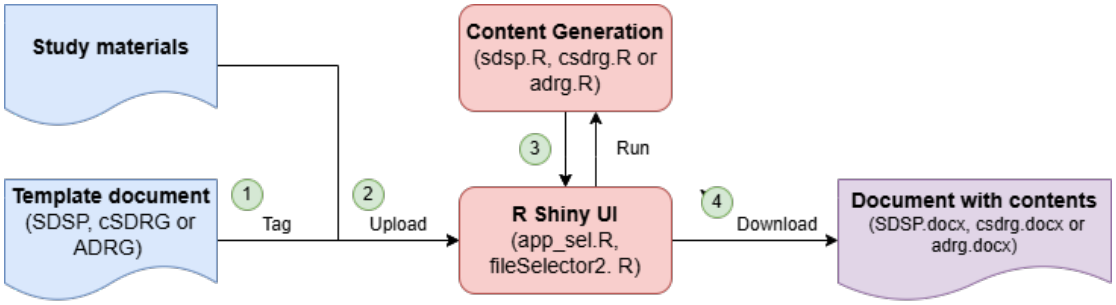

Figure 1. The workflow of SemiSPGR

1. **Template Preparation:** Standard templates for SDSP, cSDRG, and ADRG are modified to include specific placeholder tags ([start] and [end]) demarking sections where dynamic content will be inserted or existing placeholder text will be replaced. The standard templates for SDSP, cSDRG and ADRG are adapted from PHUSE Advance Hub.
2. **Template and Study Material Upload**: The user accesses the R Shiny UI and uploads the required materials for the desired document(s). Materials required are listed in **Table 2**.
3. **Automated Content Generation & Document Assembly:** After uploading required study materials, clicking a "Run" button in the UI. R scripts will execute a series of operations:
    1) Parse input files (metadata, reports, etc.).
    2) Generate specific content sections (e.g., tables of domain information, summaries of Pinnacle 21 reports).
    3) Unzip the tagged Word template and load document.xml file as XML format.
    4) Locate the [start] and [end] tags within the XML structure.
    5) Insert the R-generated content and/or replace content between these tags.
    6) Convert the modified XML back into a Word document format.
4. **Document Download:** If R scripts execute successfully, “Download” button will appear on the Shiny UI. The user can then download the populated document(s) with Microsoft Word format (.docx) directly from the Shiny UI. The SDSP named SDSP.docx. The cSDRG named csdrg.docx. The ADRG named adrg.docx.

For SDSP, the summarized tables about SDTM datasets, supplemental qualifiers, ADaM datasets and ISS/ISE datasets will be occupied. For cSDRG, the summarized tables of TA/TE/TI, SDTM subject domains including supplemental qualifiers description, issues summary from Pinnacle 21 will be occupied. For ADRG, the summarized tables of core variables, ADaM programs, analysis output programs, issues summary from Pinnacle 21 will be occupied.

TECHNICAL IMPLEMENTATION DETAILS

1. **Template Tagging:**

   The foundation of SemiSPRG's automation lies in specially prepared Word document templates. Specific locations within these templates where study-specific information needs to be inserted or updated are marked with unique tags. For instance, a section for Core variables in ADRG might be enclosed by (insert table aSection3.1 start) and (insert table aSection3.1 end) (**Figure 2**). These tags act as anchors for the R processing engine. Summary of each section tags in PHUSE template is described in **Table 1**.

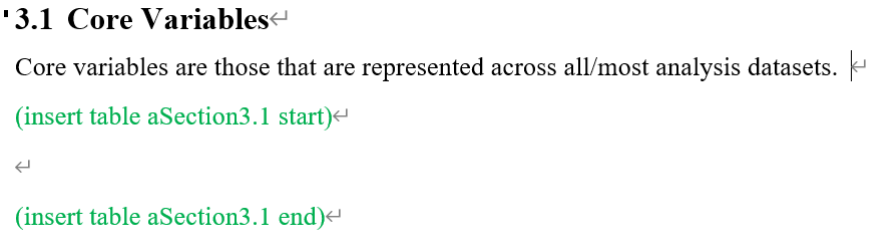

Figure 2. Tagging ADRG template document

Template	Section	Tag
SDSP template	2.SDTM Datasets	(insert table SDTM end)
	3.Supplemental Qualifiers	(insert table suppqual start) (insert table suppqual end)
	4.ADaM Datasets	(insert table ADaM start) (insert table ADaM end)
	5.ISS and ISE	(insert table ISS start) (insert table ISS end)
cSDRG template	2.3 Trial Design Datasets	(insert table TRIAL start) (insert table TRIAL end)
	2.3.1 TA – Trial Arms	(insert table TA start) (insert table TA end)
	2.3.2 TE – Trial Elements	(insert table TE start) (insert table TE end)
	2.3.3 TV – Trial Visits	(insert table TV start) (insert table TV end)
	3.4 SDTM Subject Domains	(insert table SDTM start) (insert table SDTM end)
	3.4.X XX – domain label Note: XX is domain	(insert table SUPPAE start) (insert table SUPPAE end) (insert table SUPPS start) (insert table SUPPS end) Note: this is for supplemental qualifiers description if applicable for each domain description section. The

		SUPPS represents each XX supplemental qualifiers.
	4.2 Issues Summary	(insert table P21 start) (insert table P21 end)
	Appendix I: Inclusion/Exclusion Criteria	(insert table TI start) (insert table TI end)
ADRG template	3.1 Core Variables	(insert table aSection3.1 start) (insert table aSection3.1 end)
	6.2 Issues Summary	(insert table aSection6.2 start) (insert table aSection6.2 end)
	7.1 ADaM Programs	(insert table aSection7.1 start) (insert table aSection7.1 end)
	7.2 Analysis Output Programs	(insert table aSection7.2 start) (insert table aSection7.2 end)

Table 1. Summary of each section tags

2. **Input Materials:**

The specific materials required depend on the document being generated. All materials for generation are required to be prepared and uploaded to UI. The materials needed are described in **Table 2**.

Generation	**Study Materials**	**Description**	**format**
SDSP generation （SDSP.docx）	SDSP template (tagged)	Adapted from PHUSE Advance Hub	.docx
	SDTM metadata	Describe SDTM datasets and variables information in each study. Note: if no study SDTM metadata, SDTM standard metadata will be used	.xlsx/.xlsm
	ADaM metadata	Describe ADaM datasets and variables information in each study. Note: if no study ADaM metadata, ADaM standard metadata will be used	.xlsx/.xlsm
	Supplemental Qualifiers (SUPP) variables file	Pre-described datasets and variables information about --SUPP	.xlsx
cSDRG generation （csdrg.docx）	cSDRG template (tagged)	Adapted from PHUSE Advance Hub	.docx
	SDTM metadata	Describe SDTM datasets and variables information in each study.	.xlsx/.xlsm
	SDTM datasets	Domain	.sas7bdat
	SDTM Pinnacle 21 report	SDTM Result from running Pinnacle 21	.xlsx
ADRG generation (adrg.docx)	ADRG template (tagged)	Adapted from PHUSE Advance Hub	.docx

	ADaM metadata	Describe ADaM datasets and variables information in each study.	.xlsx/.xlsm
	ADaM/TFL programs	Programs to generate ADaM and key output	.sas
	Key output	Analysis result	.pdf
	ADaM Pinnacle 21 report	ADaM Result from running Pinnacle 21	.xlsx

Note: For study SDTM/ADaM metadata, the column [STUDY_xxx] should be added in each sheet and occupied with “Y” for each row that needed.

Table 2. Study materials description

3. **R Shiny User Interface:**

The R Shiny UI provides a user-friendly platform for interaction. See in **Figure 3**. It includes:

1) Navigation bar is for the type of document to generate (SDSP, cSDRG, ADRG). The user can choose what document that they want to generate.
2) File upload widgets is for each required input material. The upload uses file selection in folder (**Figure 4**). The root directory when setting the folder path is two levels above the R Shiny code's location. This choice based on the particular way the Sanofi study folders are structured. Uploaded document format has been fixed, and is consistent with **Table 2**. Function fileInputButtonUI and fileInputButtonServer are created to facilitate file selection in R shiny UI.
3) A "Run" button is to initiate the document generation process. After clicking on “Run” button, Automated content generation will be processed. Message “Data processed finished” and a “Download” button will appear on the UI if generation process is finished and succeeded.

4) A “Download” buttons is for the generated documents after process completed (**Figure 7**). The generated documents are saved in subdirectory where UI programs are located in.

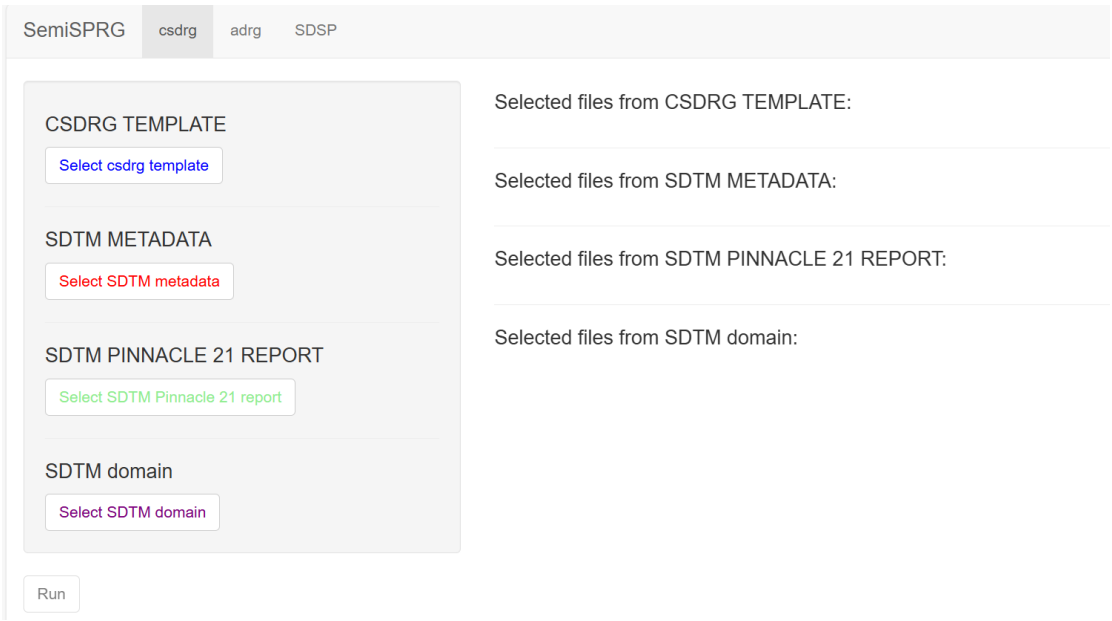

Figure 3. SemiSPRG Shiny UI

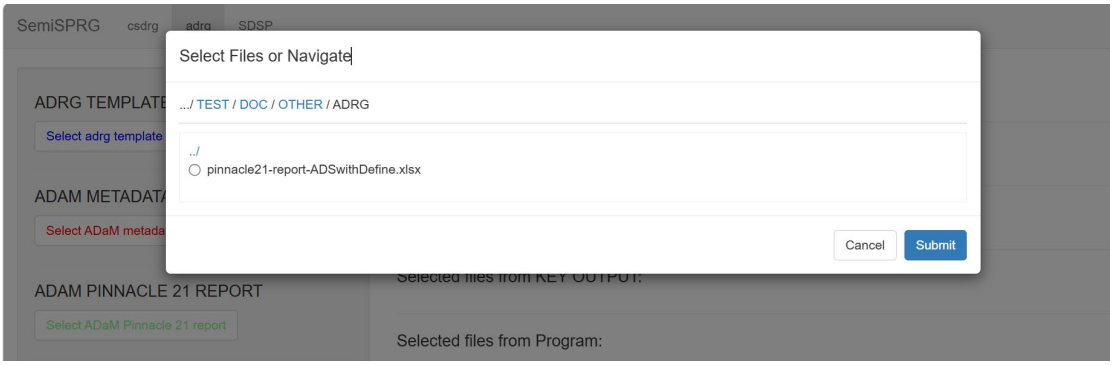

Figure 4. Pinnacle 21 report file selection in folder

4. **R-based Processing Engine:**

   The core logic resides in R scripts that perform several key tasks:

   1) Parsing Inputs: R functions read and parse the uploaded materials. For example, “readxl” package for Excel files, and text parsing for Pinnacle 21 reports.

   2) Content Generation: Based on the parsed inputs, R scripts dynamically generate the content to be inserted. “officer” and “flextable” packages are used to describe the layout of content.

   a) **SDSP:** Generates tables summarizing domain structures and SUPP dataset variables.

   b) **cSDRG:** Generates tables of SDTM domain information and incorporates key findings or summaries from the SDTM Pinnacle 21 report.

   c) **ADRG:** Generates tables of ADAM dataset information, lists of ADAM programs, TFL information, and summaries from the ADAM Pinnacle 21 report.

   3) Document Manipulation (XML Transformation):

   The tagged Word document template is internally unzipped and loaded as an Office Open XML (OOXML) format if it's a .docx file. The R script then parses this XML file to locate the [start] and [end] tags by using xml2 package. The content between these tags is replaced with the newly generated R content (which itself might be formatted as XML snippets or plain text depending on the target structure). The modified XML file is then saved back into the original Word document format.

CASE STUDY

To demonstrate the utility of SemiSPRG, we present a case study involving the generation of an ADRG for a study.

USING SEMISPRG

Prior to initiating the SemiSPRG software, it is imperative to ensure that both the tagged template and all study-specific materials, as delineated in Table 1, have been meticulously prepared. Furthermore, the R Shiny User Interface (UI) scripts, specifically app_sel.R and fileSelector2.R, are required to be located two directory levels above the primary study folder. An exemplary directory structure for a representative study, designated “ABC123”, is illustrated in **Figure 5**.

Figure 5. The folder structure of Study ABC123

Once all prerequisite materials are confirmed to be ready, the app_sel.R script should be opened within an R environment. Execution of the R Shiny UI programs is commenced by selecting the "Run App" button, typically situated in the upper-right section of the RStudio interface. This action will launch a UI, an example of which is depicted in **Figure 3**. Within this UI, the user should select the "adrg" button to facilitate the uploading of the corresponding materials. Upon successful upload, the directory path for each uploaded file will be displayed on the right-hand side of the UI, as demonstrated in **Figure 6**. After all materials have been uploaded, the content generation process is initiated by clicking the "Run" button. Successful execution is indicated by the "Run" button's label changing to "Run complete". Concurrently, a "Download" button will become active, and a confirmation message, such as "Data process finished", will appear in the UI (refer to **Figure 7**). Finally, selecting the "Download" button will result in the ADRG, named adrg.docx, being saved to the [Study_Folder]/DOC/OTHER subdirectory (e.g., ABC123/DOC/OTHER).

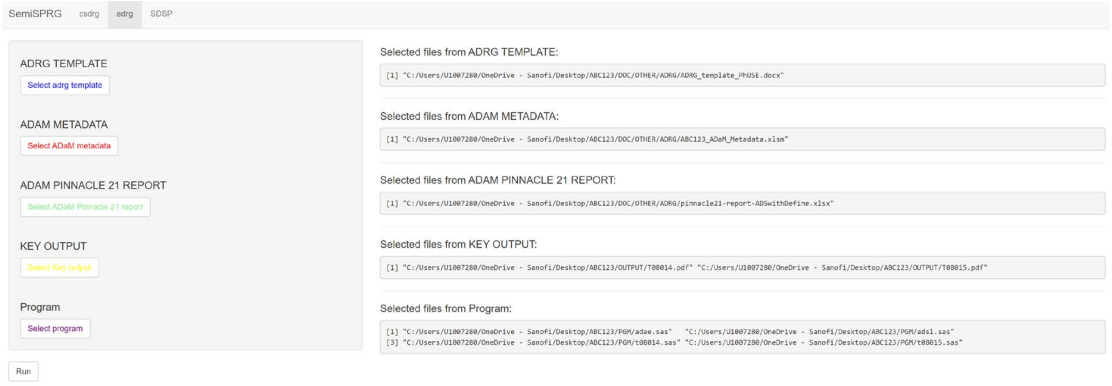

Figure 6. UI for selected files

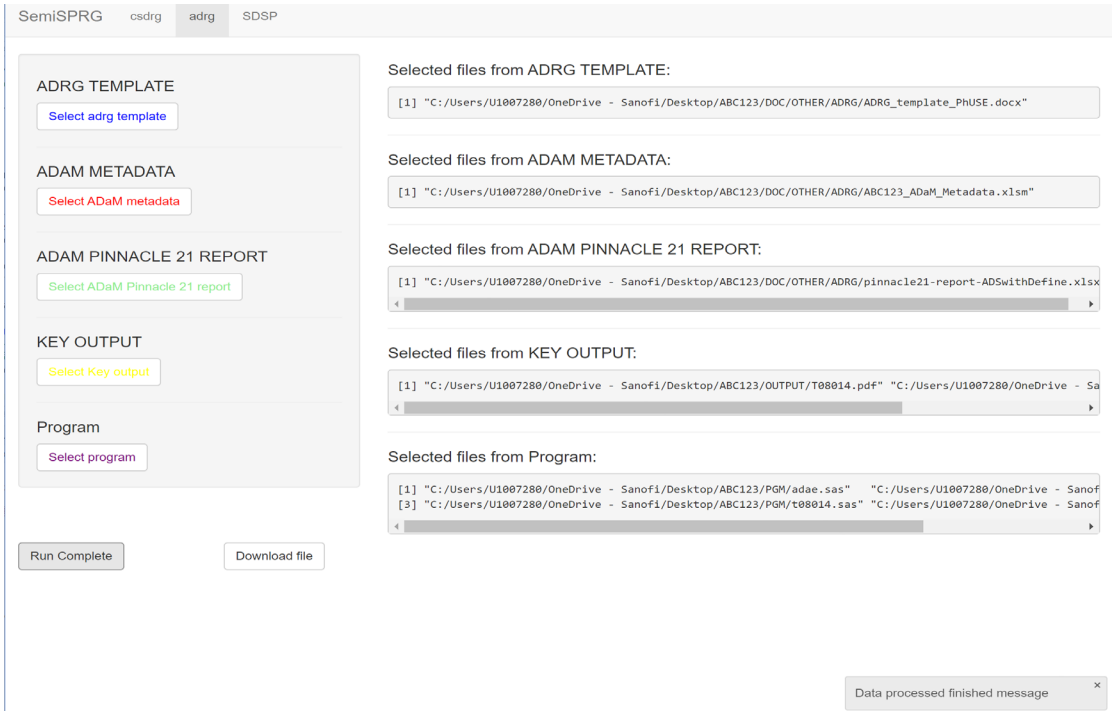

Figure 7. Run completed

RESULT

Within ADRG template, designated sections are systematically populated. Specifically, Section 3.1 is dedicated to presenting a comprehensive table that meticulously delineates the core variables extracted from the Analysis Data Set Subject Level (ADSL); an illustrative example of this tabular representation is provided in **Table 3**. Furthermore, Section 6.2 will incorporate a table summarizing the validation findings reported by Pinnacle 21. It is important to note that this summary will purposefully exclude the detailed explanation for each identified issue, as depicted in the example within **Table 4**. Subsequently, Section 7.1 provides a thorough description of the critical programs employed, encompassing both those utilized for the generation of Analysis Data Model (ADaM) datasets and the key programs responsible for producing primary analytical outputs, with a representative layout shown in **Table 5**. Finally, Section 7.2 is structured to display crucial information pertaining to these key analytical outputs themselves, offering insights into the results generated, as exemplified in **Table 6**. This methodical population ensures that the ADRG serves as a comprehensive reference document.

Variable Name	Variable Description
STUDYID	Study Identifier
USUBJID	Unique Subject Identifier
SUBJID	Subject Identifier for the Study
SITEID	Study Site Identifier

Variable Name	Variable Description
COUNTRY	Country
AGE	Age

Table 3. Example of Core Variables

Dataset(s)	Pinnacle ID	Diagnostic Message	Severity	Count	Explanation
ADAE	CT2002	RACE value not found in 'Race' extensible codelist		398	
ADSL	CT2002	RACE value not found in 'Race' extensible codelist		74	

Table 4. Example of issues summary

Program Name	Output	Macro Used
adae.txt	adae	
adsl.txt	adsl	
t08014.txt	t08014	
t08015.txt	t08015	

Table 5. Example of ADaM and TFL Programs

Program Name	Output Number	Title	Input
t08014.sas	T08014	Table 8.14: Disposition by randomized group - All Randomized Subjects	ADSL ADEX ADIS SV
t08015.sas	T08015	Table 8.15: Immunogenicity analysis sets by randomized group for infant stage - All Randomized Subjects	ADSL

Table 6. Example of Analysis Output Programs

CONCLUSION

SemiSPRG demonstrates the potential of R and Shiny to create powerful, user-friendly tools for automating critical documentation processes in clinical research. By systematically populating tagged templates with study-specific information, SemiSPRG streamlines the generation of SDSPs, cSDRGs, and ADRGs. This automation leads to significant gains in efficiency, accuracy, and consistency,

ultimately supporting higher quality regulatory submissions and more effective review of clinical study data. As the demand for rapid and reliable clinical trial reporting grows, tools like SemiSPRG will play an increasingly important role in the clinical research ecosystem.

REFERENCES

[1] Research, Center for Drug Evaluation and. “Electronic Common Technical Document (eCTD).” FDA, 4 Oct. 2024, https://www.fda.gov/drugs/electronic-regulatory-submission-and-review/electronic-common-technical-document-ectd.

[2] Wickham, Hadley. *Mastering Shiny: Build Interactive Apps, Reports, and Dashboards Powered by R*. O’Reilly Media, Incorporated, 2021.

[3] Wickham, Hadley, and Garrett Grolemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. First edition, O’Reilly, 2016.

CONTACT INFORMATION

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

Jinjing HE
Sanofi
jinjing.he@sanofi.com

Chasen CHEN
Sanofi
chasen.chen@sanofi.com