

2018 Clinical Data Landscape Survey

James McDermott, Achieve Intelligence Ltd., Oxford, UK
Guy Garrett, Achieve Intelligence Ltd., Brighton, UK
Dante Di Tommaso, Independent Consultant, Seattle, USA

ABSTRACT

In 2016 twenty of the largest BioPharma companies participated in an industry survey regarding Statistical Computer Environments. The analysis focused on user numbers, software utilization and the state of transformation projects.

In 2018 Achieve Intelligence, headed up by our chairman James McDermott, ran another survey; with an increased scope to cover all aspects of the clinical data value chain (data capture; query identification; query management; data transformation & compliance; tables, figures and listings production; statistical analysis; and data visualization). Following discussion with the contributors to the 2016 survey, we also looked to increase the breadth of participation to include a larger community.

This paper discusses the desired analysis of the survey results, and the expected benefits for the industry, and an overview of the preliminary results. With this early publication, we aim to inspire further participation and a richer industry discussion. We will follow up with select deep-dive interviews in 2019, and subsequent publications of our complete analyses.

INTRODUCTION

Achieve Intelligence has a long-term objective to understand the landscape of increasingly integrated biomedical data and replay that to the industry in order to inform and inspire best practices within our sector, and to track evolving trends and patterns.

The specific objective of the 2018 Clinical Data Landscape survey is to understand various aspects of the industry at an overview level and with further analysis broken down by specific functions along the data value chain.

DESIRED RESULTS

The 5 key questions being answered are to provide insight into the following areas:

- **Function Size** - What structures do organizations put in place to conduct these activities?
- **Business-as-Usual Effort** - What activities consume these organizational resources?
- **Technical Diversity** - What software and systems support this workforce?
- **Transformational Effort** - How are these structures and platforms evolving?
- **Vision State Solution Type** - As the industry becomes increasingly data-driven, how will and must this landscape evolve?

DATA VALUE CHAIN FUNCTIONS

These questions give an overall indication to the health of clinical data practice; with a breakdown into these seven specific functional areas.

- **Data Capture** - Electronic data capture, labs, genomics, biomarkers, imaging etc.
- **Data Query Generation** - Logical consistency checks within and between domains.
- **Data Query Management** - Workflow coordination of data query resolution.
- **Data Transformation & Compliance** - Building and verifying the transformation of source data to target data - i.e. raw to other model, such as SDTM.
- **Statistical Analysis** - Modelling, testing & estimation using statistical methodologies.
- **Tables, Figures & Listings Production** - The systematic design and publication of structured outputs in multiple formats (tables, figures, listings).

PhUSE EU Connect 2018

- **Data Visualization** - Ad-hoc or self-service analyses intended to enable users to explore and understand the significance of the data by placing it in a visual context

Fig.1: Data Value Chain Functions

Note: It is understood that not all companies adopt this assembly line approach. However, the majority do, and in the interests of consistency the participants have answered the questions as best they can within their own approach.

The following section presents the preliminary results to the five key questions with some contextual insight, followed by our concluding observations.

QUESTION 1 – COMPANY SIZE INDICATION

Within each function of the data value chain we asked a question regarding the number of user accounts on the primary system of use to perform that function. We asked participants to segment the user base into three types, illustrated with a racing car analogy.

Fig.2: User Type Definition

EXAMPLES SPECIFIED FOR EACH FUNCTION

The survey solicits the same information for each of the seven value chain functions, and provides examples tailored for each function as guidance:

Tab.1: User Type Examples per Function

Function	Admin	Producer	Consumer
Data Capture	Admin	Such as a user who builds bespoke screens	Such as a data entry operative
Data Query Generation	Admin	Such as a user who writes bespoke scripts to validate or query data	Such as a user who selects and executes pre-defined scripts
Data Query Management	Admin	Such as a user who writes bespoke scripts to send to investigator	Such as a user who manages query resolution using standard functionality
Data Transformation & Compliance	Admin	Such as a user who writes bespoke scripts to transform data	Such as a user who operates standard workflow to curate data
Statistical Analysis	Admin	Such as a user who writes bespoke algorithms to analyse data	Such as a user who operates standard workflow and built-in and tested statistical functionality
TFL Production	Admin	Such as a user who writes bespoke scripts to produce required outputs of any form	Such as a user who operates standard workflow and built-in and tested functionality to produced outputs
Data Visualization	Admin	Such as a user who writes bespoke scripts to visualize and interrogate the data	Such as a user who operates standard workflow and built-in and tested functionality to visualize and interrogate the data

The rationale for asking this question was to give an indication of the functions size within the organizations.

PhUSE EU Connect 2018

SURVEY QUESTION WORDING

What is the approximate number of user's accounts on your current primary *[DATA VALUE CHAIN FUNCTION]* system?

QUESTION CONSIDERATIONS

This measure is a proxy for the relative efforts required to administer, customize and utilize the systems for each function. Some user accounts are active daily, some occasionally or rarely, and typically to varying intensities. As often needed in complex work environments, individuals often have their own account on multiple systems. Therefore, these tallies are to full-time equivalents (FTEs). The relative counts are nonetheless informative.

Fig.3: Function Size (User Accounts)

COMMENTARY

The number of user accounts in *Data Capture* towers over those in other systems. This relates to large number of consumer users at investigator sites capturing clinical study results. The number of users that then handle queries arising from capturing those data, should raise questions of effectiveness and efficiency. The figures from Data Management, Statistical Analysis and TFL Production show effort evenly apportioned. With the growth in the discipline of Data Science it is encouraging to see healthy allocations of effort assigned to Statistical Analysis and Data Visualization, the key functions that derive value from the data.

QUESTION 2 – BUSINESS-AS-USUAL EFFORT

To understand the business-as-usual effort in supporting these functions we used the answers from question 1 to delve deeper into the Producer/Consumer split by function. Producing a box and whiskers plot enabled us to see average function sizes, with high and low outliers for each company.

SURVEY QUESTION WORDING

Note: The information to answer this question was derived from the same answered provided to Question 1.

QUESTION CONSIDERATIONS

It is understood that some users would fall into both the producer and consumer categories for this question.

Fig.4 Industry Effort by Function

COMMENTARY

Note: This Box & Whisker graph is plotted on a Log10 axis.

The first observation is that the three capture & query functions are an order of magnitude greater than all other functions, with a mean in the range of 500-1000 for the first 3, compared to a mean in the range of 200-300 for the last 4. The “boxes” show quite wide interquartile ranges, indicating significant variability between companies on the number of user accounts in each function. In summary there are significantly fewer people deriving insights from the data compared to those capturing and cleaning the data.

The proportion between consumers and producers is of the right order for the capture and query functions; however, the proportion of producers in data management and analysis & reporting functions is higher. (i.e. Capture is less than 1%, whereas, for instance, TFL Production is approximately 60%). This could indicate the sheer complexity of data management and analysis tasks, or that the software provision in this area is less mature than at the front-end of the data value chain.

QUESTION 3 – TECHNICAL DIVERSITY

We looked at Primary, Secondary & Tertiary systems, since several of the functions have multiple tools, either in combination or in parallel, to support their activities. A common scenario involves ongoing studies continuing on a mature system, with new studies starting up on a recently rolled-out system. For ease of analysis we focused on the primary system responses – further details are available to the participants regarding secondary and tertiary system use. Primary was loosely defined as the system that is used for more than 50% of your current, ongoing work.

An understanding of the software landscape is crucial for our participants so that they can ensure best practices are utilized and challenges identified for successful risk mitigation.

SURVEY QUESTION WORDING

Please indicate the software (and the approximate usage percentage) that are currently used in the area of [DATA VALUE CHAIN FUNCTION]

PhUSE EU Connect 2018

(Note: If you are part-way through a project to replace this software please answer regarding the current software - future project questions are asked later)

QUESTION CONSIDERATIONS

The main software tools in each function were listed, including “Bespoke” if the solution was hand-crafted in-house using programming languages (as opposed to software products). We also offered an opportunity for the participant to choose Other and a section for them to specify that alternative option.

Fig. 5.1 Technical Diversity [Data Capture]

Fig. 5.2 Technical Diversity [Data Query Generation]

Fig. 5.3 Technical Diversity [Data Query Management]

Fig. 5.4 Technical Diversity [Data Transformation]

PhUSE EU Connect 2018

Fig. 5.5 Technical Diversity [Statistical Analysis]

Fig. 5.6 Technical Diversity [TFL Production]

Fig. 5.7 Technical Diversity [Data Visualization]

COMMENTARY

The early half of the data value chain is dominated by the use of Medidata RAVE – it is interesting to note that this section of the process can be serviced by an established Commercial-Off-The-Shelf(COTS) solution. SAS and bespoke solutions continue to dominate transformation, analytical and reporting functions. The assumption here is that in-house solutions rely heavily on SAS software. The final upcoming section of data visualization holds a mainly 50/50 split between SAS and Spotfire. Seeing as this area is predicted to grow over the coming years data visualization vendors will be seeing this as a developing market.

QUESTION 4 – TRANSFORMATIONAL EFFORT

Many, if not all, companies are in one state of transformation or another. This question was asked in order to understand the state of the transformational effort across the industry. Are we fairly stable, or in a major upheaval of systems and infrastructure? To determine this, we asked two questions to produce a surface plot for each function.

PhUSE EU Connect 2018

SURVEY QUESTION WORDING

What is the state of your current primary system for *[DATA VALUE CHAIN FUNCTION]*

1. Current System is OK
2. Current System needs improving
3. Current System needs transforming

AND

If applicable, at what stage is your improvement project for *[DATA VALUE CHAIN FUNCTION]*

- a. Project not planned
- b. Project in planning phase
- c. Project in delivery stage
- d. Project just completed

QUESTION CONSIDERATIONS

Among participant without active projects, the survey distinguishes between no improvement undertaken (a. Project not planned) and improvement recently delivered (d. Project just completed).

Note: The depth axis has been formatted thus,

- a. Project not planned = 3-No Plan,*
- b. Project in planning phase = 2- Planned,*
- c. Project in delivery stage and d. Project just completed = 1-Good.*

Fig. 6.1 Transformational Landscape [Data Capture, Data Query Generation & Data Query Management]

Fig. 6.2 Transformational Landscape [Data Transformation & Compliance, Statistical Analysis & TFL Production]

Fig. 6.3 Technical Diversity [Data Visualization]

COMMENTARY

The axes of these plots have been arranged in order to give the most visual impact. If the plot profile is very thin (i.e. the surface is obscured) then most projects are poorly planned for systems that are in need of transformation. If, however, the plot profile is very fat (i.e. the surface is visible) this means that most projects are well underway, and very few systems need complete transformation.

From the graphs above we can see that for the first three functions in the data value chain there are several improvement projects yet to be planned, and a few transformation projects in the planning and execution phases. For the middle section of the data value chain (data transformation & compliance, statistical analysis and TFL production) we can determine there are a fair number of planned improvement and transformation projects, but not many in the execution phase. No organization is in the position that they recognize that their systems need to change and yet are doing nothing about that. Finally, the data visualization project landscape has no system needing a complete transformation (indicative of this being a fairly new area in the data value chain).

QUESTION 5 – VISION STATE SOLUTION TYPE

This question was designed to understand if the industry is moving towards one specific solution type for their desired Vision State. Vision State is not surprisingly in flux as most companies balance trade-offs between stable, validated systems and new products and technology. Data-driven organizations necessarily need cutting edge technology to both provision and process expanding data repositories and streams.

SURVEY QUESTION WORDING

If you are considering replacing your current software, please indicate the software (and the approximate usage percentage) **that will be implemented** in the area of *[DATA VALUE CHAIN FUNCTION]* (Note: this is a future project related question)

AND

If applicable, how would you describe the type of solution your improvement project will be implementing in the area of *[DATA VALUE CHAIN FUNCTION]*

Commercial Off-The-Shelf(COTS)
Configured Off-Premise
Configured On-Premise

PhUSE EU Connect 2018

Customized

QUESTION CONSIDERATIONS

Defining the types of solution was a challenge here. Commercial Off-The-Shelf(COTS) are packages which work “out of the box” with very minimal configuration. Configured is a solution that requires significant configuration for the company’s ways of working. Customized is a “from-scratch” solution that would not be appropriate to be used at another organization.

Fig. 7.1 Vision Solution Type [Data Capture]

Fig. 7.2 Vision Solution Type [Data Query Generation]

Fig. 7.3 Vision Solution Type [Data Query Management]

Fig. 7.4 Vision Solution Type [Data Transformation]

Fig. 7.5 Vision Solution Type [Statistical Analysis]

Fig. 7.6 Vision Solution Type [TFL Production]

Fig. 7.7 Vision Solution Type [Data Visualization]

COMMENTARY

The 2016 survey suggested an industry move towards Commercial-off-the-Shelf(COTS) solutions for SCE. These are vendor provided solutions created to meet a standard set of user requirements. Certain companies are looking at composite COTS solutions. These are solutions that take certain architectural building blocks provided by commercial vendors and integrate them using micro services to form a platform for clinical data processing. These can be on-premise or off-premise. A further solution type is that which has been customized to meet the needs of specific companies.

BENEFITS OF PARTICIPATION

These results are preliminary in nature. Further companies can complete the on-line survey. We will publish updated results at PhUSE US Connect 19 in Baltimore, MD.

ACKNOWLEDGMENTS

Acknowledgments go to the participating companies, listed here in alphabetical order to preserve confidentiality. Astellas, AstraZeneca, Bayer, Biogen, Boehringer Ingelheim, GlaxoSmithKline, Grünenthal GmbH, Johnson & Johnson, Novartis, Roche, Sanofi.

CONTACT INFORMATION

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

James McDermott
Achieve Intelligence Ltd.
90-92 High Street
Evesham / WR11 4EU

PhUSE EU Connect 2018

United Kingdom
Email: james.mcdermott@achieveintelligence.com
Web: www.achieveintelligence.com

Guy Garrett
Achieve Intelligence Ltd.
90-92 High Street
Evesham / WR11 4EU
United Kingdom
Email: guy.garrett@achieveintelligence.com
Web: www.achieveintelligence.com

Dante Di Tommaso
Independent Consultant
Seattle
Washington
USA
Email: dantegd@gmail.com

Brand and product names are trademarks of their respective companies.