Improvements in Data Quality Can Boost Efficiency and Reduce Development Costs: Findings from a Survey of Pharmacometric CROs

Amparo de la Peña[1], Jill Fiedler-Kelly[1], Rebecca Humphrey[1], Jeff Barrett[2]
Simulations Plus, Inc.[1] Aridhia Bioinformatics, Glasgow, UK[2]
Contact Info: amparo.delapena@simulations-plus.com

OBJECTIVES

Modern drug development, which can take up to 15 years and cost as much as $11 billion USD, relies heavily on high-quality data[1]. Recognizing the criticality of attaining quality data that is easily convertible to analysis-ready datasets, a survey was developed to obtain baseline information on data quality and data standards, largely from a CRO perspective. Recognizing: 1) that a process of curation, quality assessment and integration is required to achieve analysis-ready data and 2) that it is often a rate-limiting step for projects with modeling and simulation deliverables, we postulate that the impact of these activities on project timelines is often under-appreciated and under-estimated.

METHODS

To obtain a diverse and representative sample of Contract Research Organizations (CROs) involved in such activities, a survey on data utilization was distributed to 44 colleagues representing 32 different companies. The CROs who were invited to respond to the survey were those who offer Pharmacometrics consulting services, including data management, and whose contact information was either available online or through professional connections of the authors. The survey consisted of 11 questions, 9 of which were multiple choice, and 2 allowed for open-ended text responses. Selected multiple-choice questions allowed more than 1 option to be chosen. Responses were gathered anonymously to ensure honest feedback, while maintaining confidentiality and protecting the intellectual property of respondents and their respective companies. The survey was developed in Microsoft Forms, and the resulting data was disclosed to all survey participants.

RESULTS

Most of the 17 survey respondents create analysis-ready datasets and develop specifications for same. The majority of respondents (65%) indicated that the data they receive was almost never, or only up to 10% of the time immediately usable, meaning that it was not cleaned, defined, or appropriately formatted. The primary reasons cited for the lack of data usability were: improper formatting for the intended analysis and data quality issues such as missing data, out-of-range values, inconsistencies in units, definitions, formulas, coding, and/or relational inconsistencies like dates/times being out of order. More than 50% of respondents also indicated a lack of definition, such as no data specifications, or no define files, as the reason for the lack of data usability. With regards to the time required to clean client data to create analysis-ready datasets, respondents reported a range of time from 3 hours to 24 hours. Assuming an average data programming cost of $250/hr, the cost impact would be between $750 and $6000 per dataset.

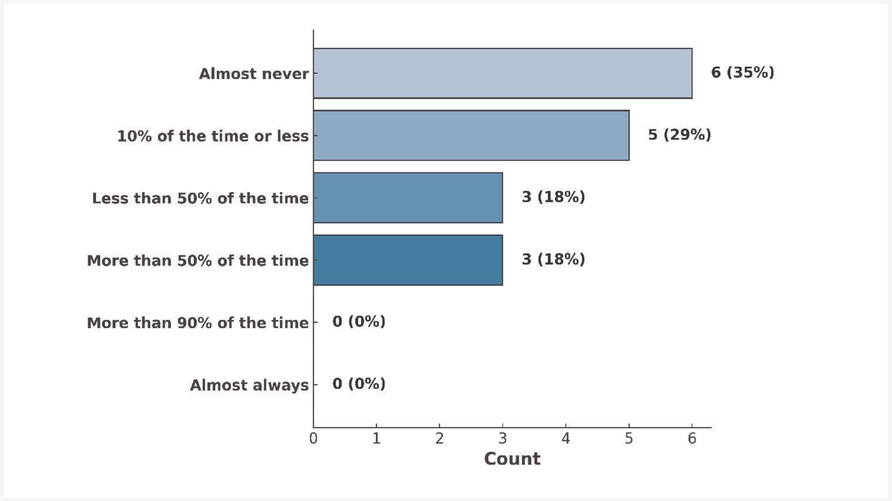

Figure 1: When receiving "analysis-ready" data from a client for work on a particular project, on average, how often was the data provided immediately usable (cleaned, defined, appropriately formatted, etc.)?

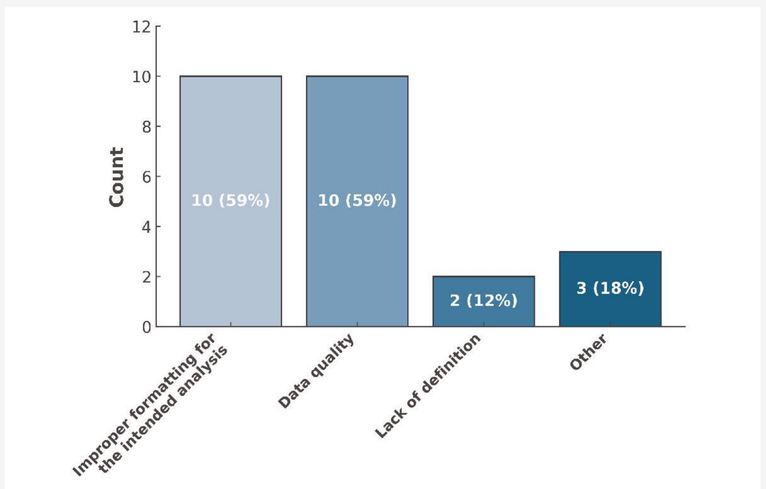

Figure 2: What is the primary reason for the lack of data usability for analysis-ready data provided by a customer?

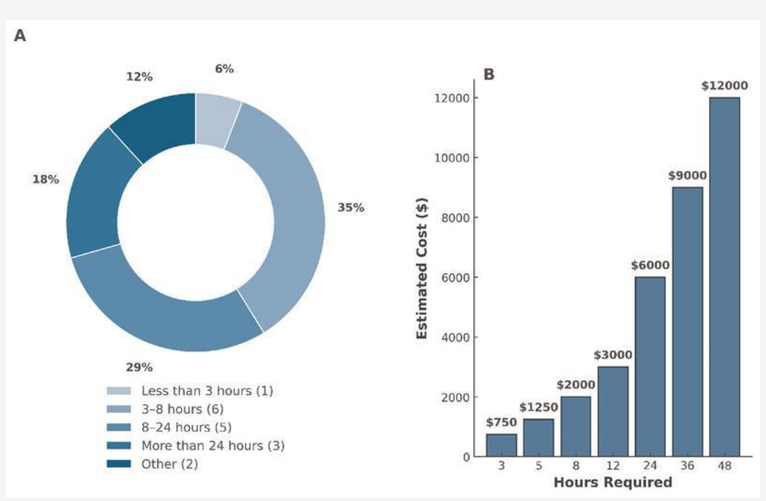

Figure 3: For datasets intended to be received as "analysis-ready": what is the typical time spent cleaning client data for improperly formatted datasets with complete data?

CONCLUSIONS

We provide herein a baseline of data quality concerns that can be ultimately linked to inefficiencies in the processing of data for analysis. Given the target audience for the survey, our findings represent the contract research organization (CRO) perspective, although we believe that the challenges identified are endemic to all sectors of the life science ecosystem (academic, industry and regulatory communities) seeking to analyze data and construct models and tools built from data, to improve drug, device and vaccine development. Our Pharmacometrics-focused CRO survey findings highlight the current practices and point to the cost of data inadequacies in both time and money, based on the effort required to curate and/or standardize the data for use in regulatory-based deliverables. Automated approaches to assessing data quality and information value would represent a milestone that further improves efficiency. Recently, multiple stakeholders have developed both open source and proprietary solutions with some predefined quality checks[2,3,4]. Automation alone cannot "fix" concerns about data quality that pervade the CRO industry and drive-up costs for Pharma and Biotech sponsors in addition to the perpetuation of delays in work completion. Recognition and acknowledgement of the current situation, as well as transparent dialogue and collaboration are necessary to improve the current state.

REFERENCES

[1] https://www.healthcatalyst.com/learn/white-papers/extended-real-world-data-the-life-science-industrys-number-one-asset

[2] https://github.com/Aridhia-Open-Source/data-quality-tool

[3] Grasela TH, et al. PAGE 19 (2010) Abstr 1901 [www.page-meeting.org/?abstract=1901]

[4] Dotan O et al. CPT PSP 12:1375–1385,2023.