Building data analytics skills for clinical drug development quality professionals - the Data Analytics University training program

Sharon Havenhand

Roche Product Ltd

Björn Koneswarakantha

F. Hoffmann-La Roche AG <https://orcid.org/0000-0003-4585-7799>

Donato Rolo

Roche Product Ltd

Pooja Mehta

Genentech Inc

Rich Bowling

Genentech Inc

Timothé Ménard (✉ timothe.menard@roche.com)

F. Hoffmann-La Roche AG <https://orcid.org/0000-0003-4545-6944>

Research Article

Keywords: Quality Assurance, Analytics, Learning, Training, Clinical Drug Development

Posted Date: February 12th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-199575/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Clinical drug development is a complex and extensive process that entails multiple stakeholders alongside patients, requires large capital expenditures and takes nearly a decade on average to complete. To ensure the correct development of this process, rigorous quality activities must be conducted to assess and guarantee the Good Clinical and Pharmacovigilance Practices (GxP) for study compliance. For about 25 years, most of these activities have been performed in the form of audits, which implies a high volume of manual work and resources in addition to being reactive by nature. Due to the limitations of this approach, together with intent to leverage new technologies in the data analytics field, a more holistic, proactive and data-driven approach needed to take place. For this to happen, quality assurance expertise needed to be complemented by the data literacy skillset. To achieve this, the Data Analytics University (DAU) was created. An in-house training program composed by two pathways that provided a framework for clinical quality staff to develop their data analytics capabilities. The first pathway covers the basics of statistics, probability and data-related terminology, while the second deepens further into the topics covered in the former followed by hands-on activities to put the knowledge to test. After successful completion of 15 DAU sessions, over 310 trained staff were able to apply their learning on data analytics and solve potential issues that might arise with a given dataset. In the near future, the DAU will be made available externally as an e-learning training program.

Background

Clinical drug development is characterised by high attrition rates, large capital expenditures and extensive timelines. Only 11.8 % of all drugs entering the clinical drug development phase will be approved, which takes 6 to 7 years and requires an investment of around 960 Million USD without accounting for the cost of failure [1]. It is a high stakes game not only for pharmaceutical companies (clinical trial sponsor) but also for patients. To identify the risk of Adverse Drug Reactions (ADRs) [2], drugs are extensively tested in animal models in the pre-clinical study phase. Drugs then enter the clinical study phase, where they are initially tested on a small sample of volunteers, and will only enter the next phase with more subjects if previous results indicate that the drug was safe. In Phase I, drugs are tested on a small sample of healthy volunteers to determine safety and dosage. In Phase II, drugs are tested on a small number of patients to have an initial reading on efficacy. Subsequent Phase III trials are larger pivotal trials on a sufficiently large number of patients to adequately prove drug safety and efficacy. A successful Phase III study will usually lead to drug approval by the competent Health Authorities (HA), but sometimes additional post-approval Phase IV studies could be requested. To assess drug efficacy and safety, appropriate data need to be collected in each phase of the clinical trials, by often hundreds of clinical trial sites that can be located all over the globe. This requires well-coordinated data management and analysis plans.

Therefore clinical trial sponsors are required by the International Conference on Harmonization (ICH) guidelines to implement and maintain Quality Assurance (QA) and quality control systems to ensure the rights, safety and well-being of research subjects and the integrity of clinical research data [3]. Following the same principles, Good Clinical Practices (GCP) and Good Pharmacovigilance Practices (GVP) have

been established as quality frameworks, the former for clinical trials, the latter for ensuring the safe use of medicinal products, once they are available to patients. Preventing harm from adverse reactions in humans arising from the use of authorized medicinal products and promoting their safe and effective use are the fundamental Pharmacovigilance (PV) objectives. Marketing Authorization Holders (MAH) are required to implement and maintain a quality system to fulfil their PV activities [4].

Quality Assurance (QA) teams of pharmaceutical companies conduct activities to assess compliance to GCP and GVP regulated activities. These activities encompass the set-up and management of a Quality Management System (QMS), including training, Standard Operating Procedures (SOPs) and Quality Strategic Activities (QSAs) such as audits. Auditing activities follow well-defined processes and have been implemented for over 25 years (at least since ICH-GCP in 1996) and involve a high volume of manual work. This is a reactive process in that audits are executed based on risk assessed from past data (from several months up to a year). For large and medium sized companies, auditing the entire “universe” of clinical trials sites on an annual basis is generally infeasible due to sheer volume, placing an even greater emphasis on a sound and timely risk assessment strategy to ensure QA activities are prioritized to assess the identified risks contemporaneously. A holistic and data-driven approach for QA that could help anticipate and reduce the risk of occurrence of key GCP/PV quality issues and that could also be used for quality by design was not available. Furthermore, it required a high level of effort and global travel by quality assurance professionals, impacting upon the work-life balance of the quality assurance professional and an ecological impact due to the high volume of transportation required to fulfill the audit schedule.

While the industry has recently been trying to leverage modern developments in data management and IT systems to facilitate the cross-analysis of clinical studies and PV processes [5,6,7,8,9,10,11], a unique skill set was needed to build and embed advanced analytics capabilities within its staff. Although clinical QA experts bring a unique skillset, “data literacy” is becoming a necessary core capability for the QA professional of the future. As a result, an in-house training program which would provide a framework for clinical quality staff to develop their analytics capabilities and increase their ability of using data-driven approaches and solutions needed to be developed: the concept of the Data Analytics University (DAU) was created.

Methods And Outline

We designed a program consisting of two consecutive pathways titled *Freshman* and *Graduate*. Targeted to quality professionals from all ages and backgrounds (see Fig. 1), the aim was to empower them to make their own self-service data requests and to perform a self-guided descriptive analysis using commonly available spreadsheet software such as Microsoft (MS) Excel. To accomplish this, the program would enable them to develop an understanding of basic data vocabulary in conjunction with a sense for statistical thinking to effectively use data products - such as dashboards - developed by quality data analysts and scientists. Among the audience on the first pathway we wanted to identify a group of quality data champions that would assist as Subject Matter Experts (SMEs) in the development of such

data products. The former pathway would establish a common ground understanding and the latter would help identify and educate data champions. (See Fig. 2)

***Freshman* pathway**

The main challenge for designing the *Freshman* pathway was the heterogeneity of the target audience. Some of them having even studied statistics (or a similar field) and others not having engaged with statistics since their high-school-level education. We needed to heavily iterate over the basic concepts to educate the latter while still keeping the former engaged and interested in order to onboard them to our *Graduate* training and identify them as possible future SMEs. We therefore initially opted for a multi-day, face-to-face, on-site training as opposed to a virtual training or e-learning. This would ensure we had enough time to cover the basics and by recruiting our data analysts/scientists along with the training designers as trainers so the more advanced participants could actively engage with them and discuss expert topics during breaks.

During the training design of the *Freshman* pathway we implemented several elements that made the training more interactive and entertaining for participants of all levels:

- a. Used relevant examples for each topic discussed on every slide, prioritizing illustrative health care examples over entertaining or abstract examples (such as astonishing scientific studies or coin flips). See Fig. 3.
- b. Introduced or closed each section covering a certain concept with an interactive poll using PollEverywhere, an online tool for conducting real-time polls within the audience to keep the engagement [12]. See Fig. 4.
- c. Every half-day session would include two or more publicly available videos of around 3-5 min length covering a data relevant topic with the help of entertaining animations.
- d. Each half-day session would be closed with an interactive knowledge check; time was given to repeat key concepts if necessary.
- e. A quarter of our training time was dedicated to a hands-on session in which participants would have to work in groups to design an analysis plan to address a relevant business problem and afterwards in the second part they would have to implement that analysis in MS Excel. The problem statement allowed the audience to design from a simple and basic to a very complex and creative solution. During these dynamic sessions we made sure that enough trainers (1 trainer per 7 participants) were present to ensure all questions could be answered.

***Graduate* pathway**

After undertaking the *Freshman* pathway, participants could apply to enroll on the *Graduate* pathway with the prerequisite to demonstrate their learning by submitting examples of how they have used data-driven approaches and solutions to help them in their daily work and how they plan to use analytics in a future project as well as taking a supplementary class on how analytics can help on the decision making. For this pathway's design, we kept the same principles as in the *Freshman*, however these classes were focussed on action-oriented problem solving skills with MS Excel (more than 50%), and the in-depth understanding of descriptive and visual analytics. The pathway concluded with the completion of a final work-related certification project.

In order to properly recognize the achievement of each participant individually and to ensure that the training met industry quality standards the program was certified by the CPD Certification Service [13], which issues online certificates that can be included in digital CVs.

Training modules

The topics covered in both pathways were developed between the data analytics/science teams together with several business SMEs and inspired by other educational formats and materials [14, 15] in conjunction with probable business problems that could be addressed using data analytics.

Freshman pathway

The contents of these modules were imparted in 2 full days divided in 4 modules of about 2 hours each.

- a. **Demystifying Analytics:** Explain common data terminology, namely - Machine Learning, Artificial Intelligence (AI), Big Data, Data Science with a focus on the health care sector.
- b. **Naked Statistics:** Descriptive Analytics on different data types in conjunction with how to calculate and interpret summary metrics.
- c. **Time to Play with Data:** Discuss a business problem at hand and address it individually using Excel.
- d. **How Data might fool you:** Understanding common data fallacies and critically interpreting data visualisations.

Graduate pathway

This pathway started with a prerequisite module that had to be completed offline, followed by 2 full days divided in 3 modules of about 2.5 hours each and with breaks in-between, ending with a final *Certification Project*.

- a. **Class 1 - Advanced Strategies for Data Analysis:** Addressed business problems using advanced MS Excel functions, namely - Pivot Tables, advanced formulas and Data Joins.
- b. **Class 2 - Data Visualisations and More:** Basic Understanding of Random Variables and Data Distributions followed by the principles of data visualisation and how to implement them in MS Excel.
- c. **Class 3 - GxP Problem Solving (with data!!):** Students were asked to put their learning into test by individually solving an analytical problem statement. This was part of the requirements needed to progress onto their Certification Project. This project would take place within the last class so students could have a chance to raise questions.
- d. **Certification Project:** On this final task students had yet a final opportunity to demonstrate their newly learned skills, this time on their own real work challenge. They were asked to propose an analytical challenge they were currently facing within their area of expertise, (e.g. GCP or GVP) and attempt to solve it by applying the knowledge acquired. This was a “take-home” project where the students would have to solve it on their own without the instructors’ help. After completion, they shared and discussed their work with the other students.

Evaluation

We performed evaluations on both pathways with two goals in mind. The first one was to give attendees a chance to check and demonstrate their knowledge and understanding; the second was to assess the impact DAU had on their learning as well as to identify any section of the program that could benefit from improvement (e.g. spend more time on a topic or to find a more illustrative example).

Freshman pathway

The exam consisted of 14 multiple choice questions. To earn a *Certificate in Data Basics* a pass mark of at least 80% was required.

As the aim of this pathway was to give a common ground of understanding, the answer to the questions were fairly straightforward. The questions focused on ensuring concepts, definitions and principles were understood. A few questions were also designed for the student to perform some basic analyses that required simple calculations.

Right after the exam’s completion, a learning transfer check was performed by running Focus Groups sessions where students formed small groups by their work areas to have an opportunity to identify and come up with real-world scenarios to put on practice what they learned. This activity encouraged them to

also do a peer-to-peer knowledge check, when bringing up ideas in the discussion, as the small number of members per group allowed everyone to actively participate.

Graduate pathway

With the purpose of assessing an individual's capability to solve a data analysis problem on their own, where there is not always a clear problem statement that indicates the start to finish path, a typical exam with a set of questions would not be enough. The best way to prove the ability to solve a real data analysis problem would be to do precisely that.

To maintain both the motivation and learning impact, we provided an initial analytical problem statement that would be challenging while at the same time achievable.

After successful completion of it, the attendees would then embark on their *Certification Project*. This time the challenge consisted in the search and successful resolution of an existent data analysis problem in their work field. The project would have to be ambitious enough so they could put in practice the lessons learned in a practical and useful deliverable in MS Excel.

Upon submission to the quality analytics team, the student would have to present their project, walk through the problem found, the reasoning and approach to the solution as well as the impact it had on the business.

Once reviewed and approved, successful students would be awarded their *Certificate* as Data Analytics University graduates.

Discussion

Impact

Freshman and *Graduate* pathways demonstrated a pronounced positive impact in the ways of working of QA professionals. Through the Data Analytics University, we were able to empower and build data literacy across QA professionals from different backgrounds. Through a review of basic statistics, data-related concepts and an in-depth session of the functionalities of MS Excel, they were able perform data-driven tasks which improved their ways of preparing and conducting quality activities.

We also saw a significant change in the kind of data-driven questions that were being put forth to the quality data analyst and quality data scientists within the team. It is important to note that QA experts bring an unmatched skill set to their role, hence it can be safe to say that they understand the technicalities and challenges of clinical study audits. When these QA experts looked at the clinical study through an analytical and data oriented perspective they were able to ask very clear questions that helped identify GCP/PV risks and mitigate them in a timely manner. The initial method of conducting audits was a reactive process and involved hand picking certain areas of investigation identified based on past

experiences around a study or a therapeutic area. This method, though tried and tested, did not encompass the entire realm of clinical data for a study. The DAU helped the QA professionals in shifting the method of audit from a reactive approach to a proactive one. After undergoing the trainings offered as a part of the DAU, the QA experts were able to use a data-driven approach to review and identify potential sites and process areas for audits. Multiple QA experts regularly started conducting risk based site selection data analyses to determine which sites pose the highest chances of GCP noncompliance for their study. Aiding to target sites based on data evidence instead of collective historical assumptions.

Towards the end of the *Graduate* pathway training, we were able to identify several data champions within the QA team who had a keen eye for data and had shown significant development in their analytical skills. Data champions participated as Subject Matter Experts within their teams and led data-driven initiatives for studies.

Results from a survey conducted to the *Freshman* pathway participants showed that most students found the program useful and applicable to their current positions, which translates into further proficiency at their daily work (see Fig. 5), e.g. audit preparation activities. In summary, the outcome of the DAU was that trained staff were able to think strategically about data and potential issues that might arise with a given dataset; to feel empowered with their new knowledge and less intimidated when working with data and to efficiently work independently with data to identify anomalies in it.

Challenges

Some of the challenges faced while conducting DAU trainings were mainly around different devices being used for example Windows vs. Mac and different versions of MS Excel. The DAU training consists of several case studies in each module. These case studies are instructor led, where the trainer goes through certain steps within Excel and the trainee must follow along. Because of different devices and different versions of software, we observed that these case studies usually took a little longer to get through as the trainees required additional assistance in following along.

A common challenge for learning programs is the *forgetting curve*, which was first theorized in 1880 by Hermann Ebbinghaus [16]. As some of the modules of the DAU were spaced in time, and as some of the trainees did not always have an immediate opportunity to apply their learning, we needed to establish a mechanism to defy the *forgetting curve*. To help retain the materials learned during each pathway, we developed a series of three short email challenges that we called *learning boosts* - the first one sent after two days, the second one sent after two weeks and the third one sent after two months once a learning pathway was completed. After the last *learning boost* set, feedback was sought to ascertain if trainees had retained the information related to the course.

The DAU trainings were also affected by COVID-19 when all the training sessions had to go completely virtual. To make sure the virtual trainings were successful, we made sure to keep the class size small (5-6 trainees per class only) so that all the trainees could be assisted and monitored effectively. Going virtual

also meant handling all technical difficulties remotely. Compared to face to face, virtual trainings were definitely more challenging. Moving forward, we plan to mitigate these issues by checking attendees software version in-advanced and provide the platform-considered instructions for each, or move to an in-browser software alternative instead.

Conclusion

In this paper, we proposed an educational program to address the existing need of upskilling quality professionals with data-related competences. We delivered a fit-for-purpose program: the Data Analytics University. An in-house designed program, it aimed to cover all relevant aspects of data analytics, while addressing the needs of quality professionals to perform their work more effectively. In consequence, successful participants no longer needed to request basic data analyses to the analytics team and upcoming requests were more advanced and analytical. Quality issues were easier identified with data (as opposed to manual review) by the end user for investigation. Data champions are now 'points of contact' that actively participate in query resolution helping reduce the analytics team burden. In the near future, the DAU program will become available to other quality professionals and HA inspectors, by offering an e-learning version of the *Freshman* pathway free to members of an independent quality assurance industry association.

Declarations

Acknowledgements

The authors would like to thank Yves Barmaz, PhD, employee of Roche/Genentech, for their significant contribution to the design, development and delivery of the Data Analytics University.

Authors Contributions

All authors wrote the manuscript. BK and TM produced the figures. BK curated and analyzed the data. DR, RB and SH quality checked the manuscript. SH, BK, TM and RB developed the Data Analytics University program.

Compliance with ethical standards

Funding

Funding was supplied by Roche/Genentech

Conflict of interest

Sharon Havenhand, Björn Koneswarakantha, Donato Rolo, Pooja Mehta, Rich Bowling and Timothé Ménard were employed by Roche/Genentech at the time this research was completed.

References

1. DiMasi AJ, et al. Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics*, 47 (2016) 20-33. <https://doi.org/10.1016/j.jhealeco.2016.01.012>
2. Guideline on Good Pharmacovigilance Practices (GVP). Module IV—pharmacovigilance audits. 2012. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-good-pharmacovigilance-practices-module-iv-pharmacovigilance-audits-superseded_en.pdf. Accessed on 11 December 2020
3. Guideline for Good Clinical Practices - International Conference of Harmonization. https://database.ich.org/sites/default/files/E6_R2_Addendum.pdf. Accessed on 11 December 2020.
4. Guideline on Good Pharmacovigilance Practices (GVP). Module I—Pharmacovigilance systems and their quality systems. 2012. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-good-pharmacovigilance-practices-module-i-pharmacovigilance-systems-their-quality-systems_en.pdf. Accessed on 11 December 2020.
5. Ménard, T., Barmaz, Y., Koneswarakantha, B. et al. Enabling Data-Driven Clinical Quality Assurance: Predicting Adverse Event Reporting in Clinical Trials Using Machine Learning. *Drug Saf* 42, 1045–1053 (2019). <https://doi.org/10.1007/s40264-019-00831-4>
6. Ménard, T., Koneswarakantha, B., Rolo, D. et al. Follow-Up on the Use of Machine Learning in Clinical Quality Assurance: Can We Detect Adverse Event Under-Reporting in Oncology Trials?. *Drug Saf* 43, 295–296 (2020). <https://doi.org/10.1007/s40264-019-00894-3>
7. Koneswarakantha, B., Ménard, T., Rolo, D. et al. Harnessing the Power of Quality Assurance Data: Can We Use Statistical Modeling for Quality Risk Assessment of Clinical Trials?. *Ther Innov Regul Sci* 54, 1227–1235 (2020). <https://doi.org/10.1007/s43441-020-00147-x>
8. Zou, M., Barmaz, Y., Prevolos, M. et al. Using Statistical Modeling for Enhanced and Flexible Pharmacovigilance Audit Risk Assessment and Planning. *Ther Innov Regul Sci* 55, 190–196 (2021). <https://doi.org/10.1007/s43441-020-00205-4>
9. Koneswarakantha, B., Barmaz, Y., Ménard, T. et al. Follow-up on the Use of Advanced Analytics for Clinical Quality Assurance: Bootstrap Resampling to Enhance Detection of Adverse Event Under-Reporting. *Drug Saf* (2020). <https://doi.org/10.1007/s40264-020-01011-5>
10. Menard, T., Bowling R., Mehta P. et al. Leveraging analytics to assure quality during the Covid-19 pandemic - The COVACTA clinical study example. *Contemporary Clinical Trials Communications* 20, 100662 (2020). <https://doi.org/10.1016/j.conctc.2020.100662>
11. Barmaz, Y., Menard T. Bayesian modeling for the detection of adverse events underreporting in clinical trials. *medRxiv* 2020.12.18.20245068 (2020). <https://doi.org/10.1101/2020.12.18.20245068>
12. Polleverywhere. <https://www.polleverywhere.com>. Accessed on 11 December 2020.
13. CPD Certification Service. <https://cpduk.co.uk>. Accessed on 11 December 2020.
14. Cairo A. *The Truthful Art: Data, Charts, and Maps for Communication*. New Riders (2016).
15. David J. Malan. What is an algorithm. <https://www.youtube.com/watch?v=6hfOvs8pY1k>. Accessed on 11 December 2020.

16. Murre JMJ., Dros J. Replication and Analysis of Ebbinghaus' Forgetting Curve. PLoS ONE (2015); 10(7): e0120644. <https://doi.org/10.1371/journal.pone.0120644>

Figures

Figure 1

(a) Age group of training participants (b) Length of year service of training participants (c) Number of Graduates per pathway

Figure 2

Data Analytics University pathways

TWO CONTINUOUS MEASUREMENTS

SCATTER PLOT

If data points are characterized by (at least) two continuous variables, we can plot them on a plane

A quick glance allows you to assess:

- ★ Statistical relationship: think of regression, correlation
- ★ Density and clusters: where are you likely to find data points?
- ★ Outliers: do some data points look odd?

GBZ patient weight and height

Figure 3

Example of training material (slide to explain scatter plots)

in WWII, a team was asked where the best place was to fit armour to a plane. The planes that came back from battle had bullet holes almost everywhere (see red spots). Where on the plane should be the best protected? Click on it

Figure 4

Example of training material (interactive poll)

Figure 5

Survey Results of 41 Graduates (a + b)