Digital Professionalism in Health and Care: Developing the Workforce, Building the Future
P. Scott et al. (Eds.)

doi:10.3233/SHTI220906

Master's Degree in Health Data Science: Implementation and Assessment After Five Years

Antoine LAMER[a,b,1] , Naima OUBENALI[c], Romaric MARCILLY[b], Mathilde FRUCHART[b] and Benjamin GUINHOUYA [a,b]
[a]*Univ. Lille, UFRS, ILIS, F-59000, Lille, France.*
[b] *Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des Technologies de santé et des Pratiques médicales, F-59000 Lille, France.*
[c]*Univ. Rennes1, LTSI, Laboratoire Traitement du Signal et de l'Image F-35000, Rennes, France.*

Abstract. Health data science is an emerging discipline that bridges computer science, statistics and health domain knowledge. This consists of taking advantage of the large volume of data, often complex, to extract information to improve decision-making. We have created a Master's degree in Health Data Science to meet the growing need for data scientists in companies and institutions. The training offers, over two years, courses covering computer science, mathematics and statistics, health and biology. With more than 60 professors and lecturers, a total of 835 hours of classes (not including the mandatory 5 months of internship per year), this curriculum has enrolled a total of 53 students today. The feedback from the students and alumni allowed us identifying new needs in terms of training, which may help us to adapt the program for the coming academic years. In particular, we will offer an additional module covering data management, from the edition of the clinical report form to the implementation of a data warehouse with an ETL process. Git and application lifecycle management will be included in programming courses or multidisciplinary projects.

Keywords. Education; Training; Curriculum; Alumni; Data Science; Healthcare

1. Introduction

The growing amount of data now available, the much more powerful computers and the new algorithms developed have made it possible to perform broader and deeper analyses than before [1]. However, the complexity of data, related to its volume, multidimensionality and poor quality, limits the traditional usage of statistics [2]. The application of mathematical models requires the automatic collection of data, whatever the source and format, the transformation, the cross-referencing of databases and the permanent storage of data to reproduce the analysis [3]. In addition to computer, mathematical and statistical skills, business domain knowledge (i.e., functional expertise) becomes crucial for understanding what "health data" really means, how those data are generated and their values. All these competencies are prerequisites for

[1] Corresponding Author, Antoine Lamer, ULR 2694, 2 place de Verdun, F-59000, Lille, France; E-mail: antoine.lamer@univ-lille.fr.

being able to translate analytical concepts and results into actionable insights for business needs, and for communicating about Artificial Intelligence (AI) or other digital health solutions to non-technical audiences and decision-makers [4].

In the healthcare and life sciences sector, the bridge discipline called Health Data Science brings together all these skills and competencies at a high level together [5,6]. Hence, health data scientists, as a human resource response to this professional and academic need, are not only data scientists in their own right, but also genuine biology-health specialists.

That is why the University of Lille, in the northern France has set up a master's degree entitled Health Data Science. In this paper, we describe the implementation of this master's degree. We assessed the program of our degree and compared it to what is required in the real business environment. For this purpose, we interviewed current students or graduates of this master's degree program to evaluate the adequacy between the proposed teaching and the needs of hiring companies and institutions.

Methods

2.1 Development of the master's degree

Before obtaining the agreement of the University of Lille in May 2018 to launch the curriculum, we conducted a one-year discussion with representatives of companies (e.g., Bayer Healthcare), government agencies (e.g., The French national digital health agency, former ASIP-Santé) and thinktanks (e.g., Healthcare Data Institute) in order to better calibrate the program to the needs of healthcare and life sciences' companies and institutions. In parallel, we carried out a comprehensive benchmarking of existing training programs in France, but also with neighboring universities in Northern Europe. The synthesis of the discussions, the benchmark, and meetings with future teachers resulted in the curriculum of the future formation. The pedagogical content of the program is summarized in Table 1.

Table 1. The Master's degree course program

Disciplines	Teaching units
Computer science	Relational databases, ontologies, programming with R and Python, Excel/VBA, big data technologies (e.g., Spark), NoSQL databases (e.g., MongoDB, Neo4j), artificial intelligence/machine learning/deep learning, bioinformatics.
Mathematics and Statistics	Algorithmics, descriptive and inferential statistical modeling, Bayesian statistics, survival analysis, spatial statistics.
Health and biology	Public health, epidemiology, molecular and systemic biology, genomics, pharmacology, physiology, physiopathology, molecular and structural microbiology, immunology.
Professional skills	Communication, professional English for data science, epistemology, training to research and scientific activities.

2.2 Assessment

Five years after the launching of the formation, we interviewed alumni or students currently in the training program. We asked them to describe (i) their academic

background before starting the program, (ii) the environments of their professional experience in the context of internship or professional contracts after the training, (iii) which aspects in their training are useful in their duties, and (iv) what are the missing skills for their practice. In addition to the pedagogical aspects, we shared with the students the impact of the COVID19 epidemic, in particular, the videoconference lessons for the students affected in 2020 and 2021.

Results

The training was initiated in 2018 and involves more than 60 teachers, 80% of them being academics and 20% coming from the industry. The degree is intended for students with a Bachelor's degree in Health and/or Life Sciences, Computer Science, Mathematics and Physics. The curriculum is also accessible to Pharmacy and Medical students, as well as people who already have other Master degrees or even Ph.D. The training takes place over two years and alternates lectures, tutorials and practical work and projects, for an annual volume of 420 hours in the first year and 415 hours in the second year. In each academic year, a mandatory 5-month internship is to take place in public institutions or private companies.

Over the 5 academic years, 53 students have already enrolled in the training and 1002 students have applied to join. More than 50 newly admitted students are expected in September 2023.

Of the 53 students who have already registered, 34 had an initial academic background in health/life sciences, 11 in computer science, and 8 in mathematics/statistics. The majority (62%) of the students started the program with a bachelor's degree, 17% already held a master's degree in another field, 15% were PharmDs or MDs, and 6% were PhDs.

3.1. Assessment

We contacted the 53 students or graduates and received 19 complete forms (36%). Among respondents, 47% were still in the program and 53% were graduated. They reported 29 professional experiences, 22 as internships and 7 as professional contracts, 14 were realized in private companies, while 15 were realized in public institutions. From students' feedback, the health courses were applied depending on the company's health sector of activity, with 31% of professional experiences held in hospitals, 28% in pharmaceutical industry, 24% in public institutions (e.g., universities), and 17,24% in private companies. Among computer science courses, they reported using or extensively using the following technologies by decreasing order: R (69%), relational databases (69%), python (50%), Excel/VBA (34%) and big data technologies (30%). Among for mathematics and statistics courses, they reported using or extensively using the following technologies by decreasing order: algorithmic (65%), inferential statistical modeling (38%), machine learning (31%), IA (27%).

3.2. Qualitative feedbacks

In addition, student respondents reported a set of lessons that they felt should be further developed so that they would feel more operational in a company or institution. For

instance, they reported the need to be initiated to and to handle the following technologies: Git and continuous integration and continuous delivery, the modeling and the implementation of data warehouses through the ETL process (Extract-Transform-Load). They also expressed the need to deepen and practice the transitions of applications from the development to the production environment, and the deployment on servers. The students also emphasized that the multi-disciplinary projects were the added value of the training and that they helped train them for their future profession.

Between November 2020 and March 2021, some courses had to be conducted by videoconference due to the COVID-19 pandemic. Students reported difficulties in following lectures, which affected the students' concentration and altered the spontaneity of the discussions. The practical courses were less impacted since they were closer to the projects that the students were used to doing outside of class.

Discussion and conclusions

We set up a master's degree in Health Data Science in 2018. After 5 years of practice, we draw up an assessment of the training, and in particular of the adequacy between the proposed teachings and the needs of companies and institutions.

Feedback from the students allowed us to identify new training needs that will help us adapt the training for the next academic year. In particular, we will offer an additional module covering data management, from editing the clinical report form to implementing a data warehouse with an ETL process. The teaching of version control, such as git, will be included in the programming courses and the lifecycle of applications, from integration and testing phases to delivery and deployment, will be covered during multidisciplinary projects.

As a bridge discipline, the teaching of Health Data Science should aim to offer lessons that involve several of the core disciplines. For this purpose, multidisciplinary projects, but also the use of data sets from the health and biology domain allow students to appreciate their future profession.

Students and alumni of the Health Data Science master's degree carry out a wide variety of missions in the healthcare and life sciences sector, and discussions with them give a good idea of what the job of a health data scientist in a company or an institution really is. The results from the interviews we conducted will be further complemented by interviews with companies' and institutions' leaders. Data science technologies and methods evolve very quickly, and it will be necessary to renew this assessment regularly to continuously adapt the training program to real needs.

The COVID-19 pandemic has left its mark on education. In particular, video-conferencing has been widely used for remote courses [11, 12]. Practical work in data science is less impacted than in other disciplines because it does not require equipment available only in the university. On the other hand, points of attention must be established in order to keep student engagement during lectures through several methods (e.g. setting expectations before class and soliciting students feedback during and after the lessons) and tools (e.g. pop quizzes, virtual whiteboards, polls).

References

[1] I.D. Dinov, Volume and Value of Big Healthcare Data, J. Med. Stat. Inform. 4 (2016) 3. doi:10.7243/2053-7662-4-3.
[2] B. Baumer, A Data Science Course for Undergraduates: Thinking with Data, (2015). doi:10.48550/arXiv.1503.05570.[3] C. Rudin, D. Dunson, R. Irizarry, H. Ji, E.B. Laber, J. Leek, T. McCormick, S. Rose, C. Schafer, M.J. Laan, L. Wasserman, and L. Xue, Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society, Undefined. (2014). https://www.semanticscholar.org/paper/Discovery-with-Data%3A-Leveraging-Statistics-with-to-Rudin-Dunson/e19ca19965b2a655f316c27cd4b45403f17a13d6 (accessed June 29, 2022).
[4] L. Cao, Data Science: A Comprehensive Overview, ACM Comput. Surv. 50 (2018) 1–42. doi:10.1145/3076253.[5] I.D. Dinov, Quant Data Science meets Dexterous Artistry, Int. J. Data Sci. Anal. 7 (2019) 81–86. doi:10.1007/s41060-018-0138-6.[6] P. Kubben, M. Dumontier, and A. Dekker, eds., Fundamentals of Clinical Data Science, Springer, Cham (CH), 2019. http://www.ncbi.nlm.nih.gov/books/NBK543527/ (accessed January 15, 2022).
[7] D. Schuff, Data Science for All: A University-Wide Course in Data Literacy, in: A.V. Deokar, A. Gupta, L.S. Iyer, and M.C. Jones (Eds.), Anal. Data Sci. Adv. Res. Pedagogy, Springer International Publishing, Cham, 2018: pp. 281–297. doi:10.1007/978-3-319-58097-5_20.
[8] R.S. Robeva, J.R. Jungck, and L.J. Gross, Changing the Nature of Quantitative Biology Education: Data Science as a Driver, Bull. Math. Biol. 82 (2020) 127. doi:10.1007/s11538-020-00785-0.
[9] O. DeMasi, A. Paxton, and K. Koy, Ad hoc efforts for advancing data science education, PLoS Comput. Biol. 16 (2020) e1007695. doi:10.1371/journal.pcbi.1007695.
[10] T.H. Davenport, and D.J. Patil, Data Scientist: The Sexiest Job of the 21st Century, Harv. Bus. Rev. (2012). https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century (accessed June 28, 2022).
[11] S. Kaup, R. Jain, S. Shivalli, S. Pandey, and S. Kaup, Sustaining academics during COVID-19 pandemic: The role of online teaching-learning, Indian J. Ophthalmol. 68 (2020) 1220–1221. doi:10.4103/ijo.IJO_1241_20.
[12] J.W. Redinger, P.B. Cornia, and T.J. Albert, Teaching During a Pandemic, J. Grad. Med. Educ. 12 (2020) 403–405. doi:10.4300/JGME-D-20-00241.1.