Assessing Health Data Science Internships: A Comprehensive Study at the University of Lille.

C. Andries, I. Ouddarour, C. Saint-Dizier, B. Guinhouya, Antoine Lamer

Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240706

Assessing Health Data Science Internships: A Comprehensive Study at the University of Lille

Coline ANDRIES[a,b,1] , Ikram OUDDAROUR[a,b], Chloé SAINT-DIZIER[a,b,c], Benjamin GUINHOUYA[a,c] and Antoine LAMER[a,b,c]
[a] *Univ. Lille, UFR3S - Ingénierie et Management de la Santé, 59000 Lille, France*
[b] *Fédération régionale de recherche en psychiatrie et santé mentale - F2RSM Psy, Hauts-de-France, Saint-André-Lez-Lille, France*
[c] *Univ. Lille, CHU de Lille, ULR 2694 - METRICS: évaluation des technologies de santé et des pratiques médicales, 59000 Lille, France*

Abstract. Data Science emerged as a new cross-disciplinary discipline at the intersection of statistics, computer science, and expertise in a specific domain, such as health and biology. The data science field, alongside other data-related professions, is continuously evolving. We conducted a study examining tasks assigned to first-year internship students pursuing a Master's degree in Health Data Science, exploring the missions, technologies employed and skills required, and internship alignment with students' training through semi-structured interviews with 32 participants. Three quarters of the students were placed in teams within the public sector. Among these entities, there were 11 hospitals and 12 universities. Although the majority of students did their internship as part of a methodological team, they often had a healthcare professional on their team. Nearly half of the missions involved descriptive analysis, followed by 9 missions focused on etiology or prediction and 8 missions on implementing a data warehouse. The majority of students had to perform data management and produce graphs, while only half conducted statistical analysis. The findings highlighted that data management remains a major challenge, and it should be taken into consideration when designing training programs. In future, it remains to determine whether this trend will continue with second-year students or if, with experience, they are more often assigned statistical analyses.

Keywords. Education; Curriculum; Data Science; Health Informatics

1. Introduction

Data Science emerged as a new cross-disciplinary discipline in the last 20th century, at the crossroads of statistics, computer science and knowledge of a specific business field, in this case health and biology [1]. Process requires skills in collection, processing, visualization, modeling, and interpretation of large quantities of heterogeneous data [2]. This results in a need for specific skills training, either through the adaptation of existing courses [3] or the creation of new ones [4]. Data science training courses in

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

healthcare address all levels, for example undergraduate medical students [5] or Master Degree [6]. In addition to initial training, these qualifications may target various profiles, including biomedical engineers and researchers [7,8], medical students [9], and nurses [10].

The data science field, alongside other data-related professions, is continuously evolving due to the emergence of emerging technologies, innovative methodologies, and evolving demands. As a result, new roles have appeared to work across the entire data exploitation and value chain, such as data stewards, data engineers, data analysts, chief data officers, and business intelligence developers. It would be worth checking out what data scientists are doing in the field to see if the training is still relevant or should be adapted.

In this study, we evaluated the tasks assigned to students pursuing a Master's degree in Health Data Science during their first-year internship. We interviewed students, gathering insights into their internship experiences, the nature of their assignments, the tools and data utilized, and details about the companies that hired them and the work environments they encountered.

2. Methods

2.1. Master Degree in Health Data Science

Initiated in 2018, the training program involves a faculty of over 60 instructors, comprising 80% from academia and 20% from the industry [6]. It caters to individuals holding 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 those with other Master's or Ph.D. qualifications. The program spans two years, incorporating a mix of lectures, tutorials, practical work, and projects, totaling 420 hours in the first year and 415 hours in the second year. A mandatory 5-month internship in either public institutions or private companies is required each academic year. In September 2023, 32 new students were admitted to the first year.

2.2. Interview

We conducted preliminary semi-structured interviews with 10 students. The interviews were conducted by telephone or face-to-face and were used to define the scope of the internships carried out in the first year. Based on these interviews, we have set up an interview grid with an eCRF [11]. It contains the following sections: (i) the internship's host company and its field, (ii) the profiles of the internship supervisor and the skills present in the company, (iii) the scope of the mission, characteristics of the data, and technologies employed, (iv) the nature of the tasks (e.g. data management, statistics, data visualization), (v) and the student's perception regarding their internship and its alignment with their training. The remaining 29 interviews were conducted via video conferencing. Finally, all the student of the degree had answered the survey (n = 39).

3. Results

Three quarters of the students were placed in teams within the public sector (n = 27, 84.4%). Among these entities, there were 11 hospitals, 12 universities, 3 public interest consortium and one institute. Only 5 students completed their internship in a private company (9.4%). The majority of students conducted their internship in a methodological team (n = 25, 78.1%), while 10 were in a medical department or another applied field (31.2%), with 3 of them being in both fields simultaneously. Approximately one-quarter of the students were supervised by a data scientist (n = 10, 30.3%) or a healthcare professional (n = 10, 30.3%), followed by a researcher (n = 6, 18.8%). In other cases, the students were supervised by a computer scientist, a statistician or another profile (Figure 1). Their team included a healthcare professional in half the cases (n = 19, 59.4%), as well as a computer scientist (n = 15, 53.1%), data scientist (n = 15, 46.9%) or researcher (n = 15, 46.9%) (Figure 1).

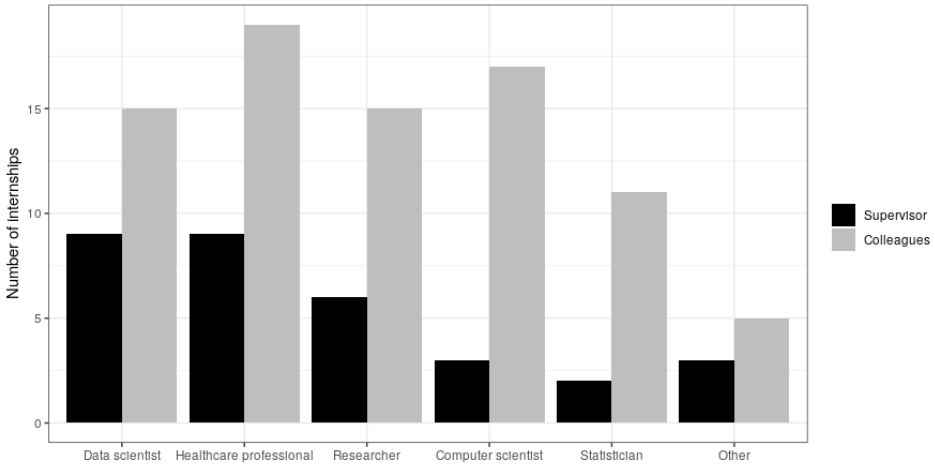

Figure 1. Profiles of internship's supervisors and colleagues

Nearly half of the missions involved descriptive analysis (n = 15, 46.9%), followed by 9 studies focused on etiology or prediction (28.1%), a quarter focused on implementing a data warehouse (n = 8), 4 involved developing a visualization tool or data collection tool (12.5%), and 3 were dedicated to improving an existing process or evaluating it (n = 3, 9.3%). The majority of students had to perform data management (n = 31, 96.9%) and produce graphs (n = 27, 84.4%), while only half of them conducted statistical analysis (n = 15, 50.0%).

Most students (96.9%, n = 31) worked with structured data, while only 8 students (25.0%) had to handle non-structured data. The main types of data used were patient data (n = 18, 56.2%), administrative data (n = 9, 28.1%), demographic data (n = 9, 28.1%), and drug delivery data (n = 7, 21.9%). They most often covered the scale of a single institution (n = 10, 31.2%), or a region (n = 10, 31.2%). Otherwise, they consisted of international data (n = 8, 25.0%), data from a population sample (n = 7, 21.9%), national data (n = 6, 18.8%), or data from other scopes (n = 4, 12.5%).

The programming languages or technologies used were mainly R (n = 24, 75.0%), SQL (n = 12, 37.5%), and Python (n = 9, 28.1%). To note, 5 have still used spreadsheet

tool (15.6%) (Figure 2). The internship deliverable took the form of a report (n = 14, 43.8%), a dashboard (n = 11, 34.4%), a presentation (n = 11, 34.4%), a database (n = 8, 25.0%), or a scientific article (n = 7, 21.9%) (Figure 2).

Twenty-five students (78.1%) reported that the internship met their expectations, while 3 students perceived it to be more focused on data management, and 3 others on other types of tasks. All students mentioned the value of the data management courses. Additionally, half of the students (n = 18, 56.2%) highlighted the contribution of health-related courses in their training, which helped them better understand the environment and the subject matter of the internship.

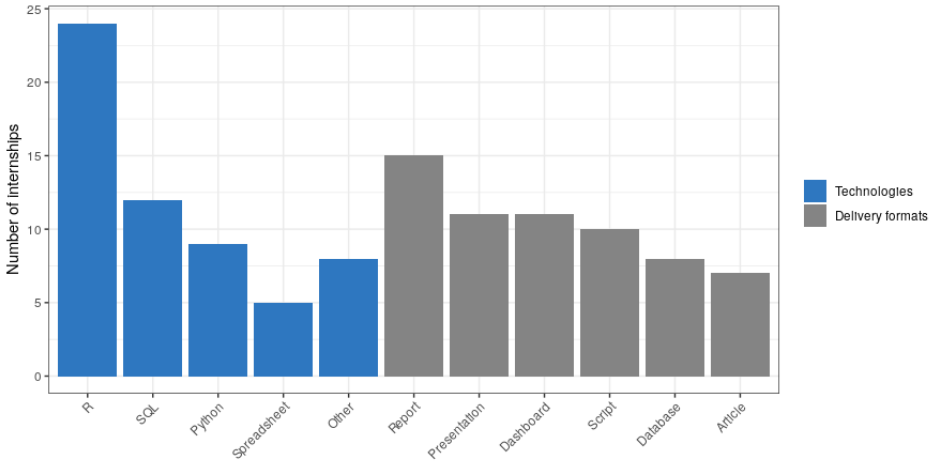

Figure 2. Technologies and delivery formats

4. Discussion and Conclusions

This study was an opportunity to gain an overview of the missions of health data scientists. It is important to have this feedback to be able to adapt the training programs as effectively as possible. The results highlighted the significance of data management and programming skills, particularly with R and SQL. Descriptive statistics and data warehousing were also identified as very frequent assignments. Conversely, there were few etiological studies conducted, and only half of the students performed statistical analyses. For first-year internships, they primarily took place in public institutions, namely universities or hospitals.

The importance of data management and data warehousing could be attributed to the substantial workload of data transformation required before proceeding to data analysis or the fact that some teams are still in the early stages of handling data. Data preparation remains a major challenge, and it should be taken into consideration when designing training programs. It has already been widely reported that data quality influences results [12]. Without proper data management and data quality operations, it becomes challenging to explore potential applications of data mining and machine learning concepts [13]. In response to this, the role of a data engineer has complemented that of a data scientist [14].

In future, it remains to determine whether this trend will continue with second-year students and whether, with more experience, they will be more frequently entrusted with statistical analysis and artificial intelligence-based technologies. This will also be an opportunity to identify new teaching needs and adapt the curriculum.

References

[1] Blei DM, Smyth P. Science and data science. Proc Natl Acad Sci U S A. 2017 Aug 15;114(33):8689–92.
[2] McDavid A, Corbett AM, Dutra JL, Straw AG, Topham DJ, Pryhuber GS, et al. Eight practices for data management to enable team data science. Journal of Clinical and Translational Science. 2021 Jan;5(1):e14.
[3] Robeva RS, Jungck JR, Gross LJ. Changing the Nature of Quantitative Biology Education: Data Science as a Driver. Bull Math Biol. 2020 Sep 19;82(10):127.
[4] DeMasi O, Paxton A, Koy K. Ad hoc efforts for advancing data science education. PLoS Comput Biol. 2020 May 7;16(5):e1007695.
[5] Baumer B. A Data Science Course for Undergraduates: Thinking with Data [Internet]. arXiv; 2015 [cited 2022 Jun 28]. Available from: http://arxiv.org/abs/1503.05570
[6] Lamer A, Oubenali N, Marcilly R, Fruchart M, Guinhouya B. Master's Degree in Health Data Science: Implementation and Assessment After Five Years. Stud Health Technol Inform. 2022 Aug 31;298:51–5.
[7] Van Horn JD, Fierro L, Kamdar J, Gordon J, Stewart C, Bhattrai A, et al. Democratizing data science through data science training. Pac Symp Biocomput. 2018;23:292–303.
[8] Dunn MC, Bourne PE. Building the biomedical data science workforce. PLoS Biol. 2017 Jul 17;15(7):e2003082.
[9] Wiggins WF, Caton MT, Magudia K, Glomski SHA, George E, Rosenthal MH, et al. Preparing Radiologists to Lead in the Era of Artificial Intelligence: Designing and Implementing a Focused Data Science Pathway for Senior Radiology Residents. Radiol Artif Intell. 2020 Nov;2(6):e200057.
[10] Dreisbach C, Koleck TA. The State of Data Science in Genomic Nursing. Biol Res Nurs. 2020 Jul;22(3):309–18.
[11] Martignene N, Amad A, Bellet J, Tabareau J, D'Hondt F, Fovet T, et al. Goupile: A New Paradigm for the Development and Implementation of Clinical Report Forms. Stud Health Technol Inform. 2022 May 25;294:540–4.
[12] Kilkenny MF, Robinson KM. Data quality: "Garbage in – garbage out." HIM J. 2018 Sep 1;47(3):103–5.
[13] Mottaghy FM, Hertel F, Beheshti M. Will we successfully avoid the garbage in garbage out problem in imaging data mining? An overview on current concepts and future directions in molecular imaging. Methods. 2021 Apr 1;188:1–3.
[14] Coursera [Internet]. 2023 [cited 2024 Mar 18]. What Is a Data Engineer?: A Guide to This In-Demand Career. Available from: https://www.coursera.org/in/articles/what-does-a-data-engineer-do-and-how-do-i-become-one