Integrating data science principles into a scientific writing course for English learners at Nippon Medical School

James Ellinger

日本医科大学基礎科学紀要　第 53 号　抜刷
令和 7 年 1 月 31 日
編集　日本医科大学基礎科学紀要編集委員会
発行　日本医科大学

〈Research Note〉

Integrating data science principles into a scientific writing course for English learners at Nippon Medical School

James Ellinger *

Introduction

This article will briefly discuss the use of data science in medicine, summarize the curriculum for data science at Nippon Medical School (NMS), and then describe how principles of data science might be integrated in a scientific English writing course. The field of data science can be defined as "an interdisciplinary academic field that uses statistics, scientific computing, scientific methods, processing, scientific visualization, algorithms and systems to extract or extrapolate knowledge and insights from potentially noisy, structured, or unstructured data" ('Data science', 2024). Research in this field often makes use of large datasets, machine learning, and artificial intelligence (AI).

The intersection of data science and medicine is expected to lead directly to advancements in personalized and precision medicine (Fröhlich *et al.*, 2018; Hulsen *et al.*, 2019), as well as early detection and diagnosis of diseases (Siuly and Zhang, 2016; Iregbu *et al.*, 2022). Additionally, integrating data science into healthcare is expected to aid in more general tasks such as electronic health records management and privacy and fraud detection (Subrahmanya *et al.*, 2022). Furthermore, as the volume of data collected from clinical trials

* 日本医科大学・外国語教室
Department of Foreign Languages, Nippon Medical School

increases, it can be expected that data science-based approaches will play an important role in driving evidence-based medicine practices (Subbiah, 2023).

Despite the buzz around data science and the potential for transforming medicine, a review on the use of big data and data science in medical education noted that current curricula and syllabi might not be adequately preparing future healthcare professionals (Khamisy-Farah *et al.*, 2021). For example, in a study that investigated practicing medical oncologists' genomic literacy, an area that includes processing large genomic datasets, nearly 50% indicated they had little knowledge of newer techniques currently in use, and more than 40% of those surveyed did not feel that genomic training offered in medical programs is adequate (Chow-White, Ha and Laskin, 2017). Another study, which focused on newly graduated medical doctors as well as those younger than 35, found that knowledge of clinical applications for data science-related areas such as big data, predictive models, and AI, was reported to be 39%, 30%, and 43%, respectively. Furthermore, for each of those three areas, only 13% responded that they had experience with direct involvement in daily clinical and research activities (Casà *et al.*, 2021). While these studies are limited in scope, they indicate that improvements in medical school curricula might be needed to prepare future doctors to incorporate data science-based approaches into their practice.

Data science at Nippon Medical School

To prepare its doctors for a future in which data science is likely to be incorporated into medicine, Nippon Medical School (NMS) established the Center for Mathematics, Data Science, and Artificial Intelligence in August of 2021. The center was established following the guidelines for the Program for Mathematics, Data Science, and AI Smart Higher Education Program (MDASH) put forth by Japan's Ministry of Education, Culture, Sports, Science and Technology (MEXT). The program at NMS has been certified at the Literacy Plus level and the overarching goals are to introduce students to the relationships between data and society. Specifically, students will learn about

the connections between medicine, AI, and data science. Unique features of Nippon Medical School's program include the use of real medical data and lectures provided by medical doctors to demonstrate the relationships to students. Students are issued a certificate of completion upon finishing the program.

The program is administered during the first 4 years of coursework as part of the Longitudinal Program（縦断型プログラム）. Regarding AI, the program covers the foundations, history, societal impacts, and applications across various fields. These fields are not limited to medicine but also include autonomous systems and data-driven industries. Key topics include machine learning, deep learning, AI ethics, and practical Python programming for tasks such as data analysis, data visualization, simulation, and image processing.

With respect to data science, this course addresses foundational skills in information security, data management, and statistical analysis, including password safety, information retrieval, and the use of public datasets. Topics also include data cleansing, questionnaire design, and visualization techniques such as histograms and boxplots. Practical skills in software applications for word processing, spreadsheets, databases, and presentations are highlighted, alongside an introduction to linear algebra and Python for tasks such as matrix operations, eigenvalue analysis, and statistical modeling.

The mathematics portion focuses on calculus fundamentals and their applications in the natural sciences, probability concepts and distributions, and statistical methods, including interval estimation for population means, variances, and proportions.

A full list of the 35 learning objectives associated with the program can be found at the center's homepage (the URL is provided in the Notes of Table 1). The items that are most relevant or which have potential relevance to the scientific writing course for English learners described in this paper are listed in Table 1. Some of these items are currently being addressed to various degrees in the writing course, while addition of others is under consideration. The details will be described further in the next section.

Table 1. Selected learning objectives from the NMS Center for Mathematics, Data Science, and Artificial Intelligence with relevance to a scientific English writing course taught at NMS.

Item	Students will be able to
1	explain why data-driven thinking is important.
5	explain the different types of data and the concepts of primary and secondary data.
6	perform basic programming in Python.
16	perform basic operations of applications like word processing, spreadsheets, and presentations.
18	summarize data using spreadsheet software and visualize it using tables and graphs.
19	create basic graphs such as histograms, bar graphs, line graphs, and scatter plots.
20	perform simple data screening using spreadsheet software.
21	use information retrieval services to conduct appropriate information searches.
29	explain discrete and continuous random variables, probability density functions, expected values, and variance.
30	explain various probability distributions (binomial, Poisson, normal, exponential, etc.).
31	explain the law of large numbers and the central limit theorem.
32	explain various methods for interval estimation of population means, population proportions, population variances, etc.
33	appropriately use different methods of data representation (visual representations such as histograms and boxplots, and numerical representations like means and variances).
34	explain the differences between scatter plots, correlation coefficients, and regression lines for two variables.
35	explain concepts like correlation versus causation, spurious correlation, and latent variables.

Notes: The full list of objectives is available at the center's homepage:
https://sites.google.com/nms.ac.jp/ai-edu/ 教育プログラム
The translations for the objectives are provided by the author and are unofficial.

Scientific writing for English learners at Nippon Medical School

Although separate from the Center for Mathematics, Data Science, and Artificial Intelligence, the Department of Foreign Languages at NMS offers English coursework in which students can put to practical use the data science skills they acquire. Some of the important aspects of data science training and education mentioned above are analysis, visualization, and communication of data. This section will describe current and possible future ways in which data science skills are applied in an English class at NMS. Linking the ideas between classes and communicating those points to students is important because it should help students consolidate their understanding of the concepts and hopefully improve motivation by demonstrating the interdisciplinary nature of their coursework.

All first-year students at Nippon Medical School are enrolled in a scientific English writing course. This course spans all three academic terms, and the main goal is for students to write an original research paper, in English, following the conventional Introduction, Methods, Results, and Discussion (IMRaD) format. Additionally, students give an oral presentation of their project in English with their classmates and teachers serving as the audience[1].

One of the defining features of this course is that students work in teams and develop a research question for which they do not know the answer. Then they design and conduct an original experiment, which is the driving force for the paper that they write. This approach is taken because it allows students to have ownership over their project. In the instructors' experience, it has been observed that this ownership provides strong motivation for the students to become engaged in the project and promotes greater curiosity than simply assigning students an experiment with given data and a known outcome.

---

[1] This course is based on the instructors' prior experience teaching in the Active Learning of English for Science Students program at the University of Tokyo (https://eng2.c.u-tokyo.ac.jp/ale_web/aless-en/).

This course is taught in a student-centered, active learning style, which means that during class time, the teachers do not lecture but instead provide students with goal-driven activities. The activities are provided in English, and unless otherwise specified, students are expected to complete the activities using English. For example, students read a published research paper, which is used as the basis for their own project. Outside of class time, students are introduced to the structure and language of a research paper by watching videos created by the instructors. Then, during class time, using guided activities created by the instructors, the students work with their team to collaboratively identify the limitations of the previous research, formulate a research question, and design an experiment to answer the question.

As mentioned above, some aspects of this course are relevant to the learning objectives stated by the Center for Mathematics, Data Science, and Artificial Intelligence (Table 1). The most immediate learning objectives that are relevant and which are directly covered in the scientific writing course are items 16 – 20. Students write their entire paper using Google Docs, gather and curate data using Google Sheets (analysis is performed with different software as described below), and prepare their oral presentation using Google Slides. These are important tools that the students use to communicate and visualize their data.

Items 33 – 35 are relevant to the analysis of data and become applicable when students prepare the Results and Discussion sections of their paper. The instructors directly address the objectives of item 33 (*appropriately use different methods of data visualization*) through activities conducted in class. However, items 34 (*explain the differences between scatter plots, correlation coefficients, and regression lines for two variables*) and 35 (*explain concepts like correlation versus causation, spurious correlation, and latent variables*) are not directly addressed in the scientific English writing course. This could be remedied in the future by having discussions with instructors who teach in the Center for Mathematics, Data Science, and Artificial Intelligence.

Currently item 21 (*use information retrieval services to conduct appropriate information searches*) is not directly addressed in the scientific writing course

since the students are provided published literature by the instructors. The instructors would like to integrate this topic into the course, but due to time constraints have been unable to do so thus far. Occasionally a student will request guidance on finding additional literature, but this is handled on a case-by-case basis. However, the instructors have observed that some students take initiative to seek published literature on their own. Coordination with members of the Center for Mathematics, Data Science, and Artificial Intelligence as well as administrators of the Longitudinal Program[2] will help the instructors address item 21 more thoroughly in the future.

Item 1 (*explain why data-driven thinking is important*) is loosely addressed via an activity that students do on the first day of class. Students are provided with a series of graphs that show the use of English in published scientific research, and they are guided through a discussion based on this data. The last question of the discussion asks students to consider why scientific research is important for medical doctors. The goals of the discussion include getting students to think about the importance of using data when answering questions as well as the role of data in medicine.

Item 5 (*explain the different types of data and the concepts of primary and secondary data*) is addressed in passing, at least with respect to concept of primary data. Students design and conduct an original experiment to gather their own data. Therefore, they gain first-hand experience with the generation of primary data. However, having the students link this back to concepts they have learning in the Center for Mathematics, Data Science, and Artificial Intelligence is not a point that has been emphasized thus far.

With regards to items 29 – 32, which involve specific statistical concepts and procedures, these are not directly addressed in the scientific writing course as the details are outside the scope of an English course. However, they are relevant to data analysis. Students need to calculate sample means and

---

[2] All students in the second year take a course called Scientific Research 2（科学的研究2）during which they learn how to use PubMed to search for scientific literature.

variances based on the data they collect. Furthermore, students receive some instruction for conducting statistical tests such as a t-test or ANOVA. The objectives in items 29 – 32 lay down some of the foundations for analyses that students might use for their own data analysis. Students will receive more direct instruction in statistics during their second year of coursework at the Center for Mathematics, Data Science, and Artificial Intelligence. However, the instructors of the scientific writing course have not yet highlighted the connections between the future statistics coursework and the scientific writing course. Doing so might help improve student motivation and consolidation of their acquired skills.

Item 6 (*perform basic programming in Python*) is relevant to the scientific writing class through its connection to items 19 (*create basic graphs such as histograms, bar graphs, line graphs, and scatter plots*) and 33. Through coursework in the Center for Mathematics, Data Science, and Artificial Intelligence students learn how to make different types of graphs using Python. Python can also be used to conduct statistical tests. Currently, in the scientific writing course with regards to data analysis and visualization, students are instructed in the use of a software application called Data Explorer[3], which is provided by the Howard Hughes Medical Institute. Data Explorer has a straightforward and easy-to-use user interface (Figure 1), but students are limited to using graph types and statistical tests that are available within the software. Using Python might allow students to take a more flexible approach to their data analysis while further building their data science analysis skills. Further discussion with members of the Center for Mathematics, Data Science, and Artificial Intelligence will allow the instructors to determine if there is a practical way to coordinate the timing and use of Python in the scientific English writing course.

---

[3] https://media.hhmi.org/biointeractive/tool/dataexplorer/v1

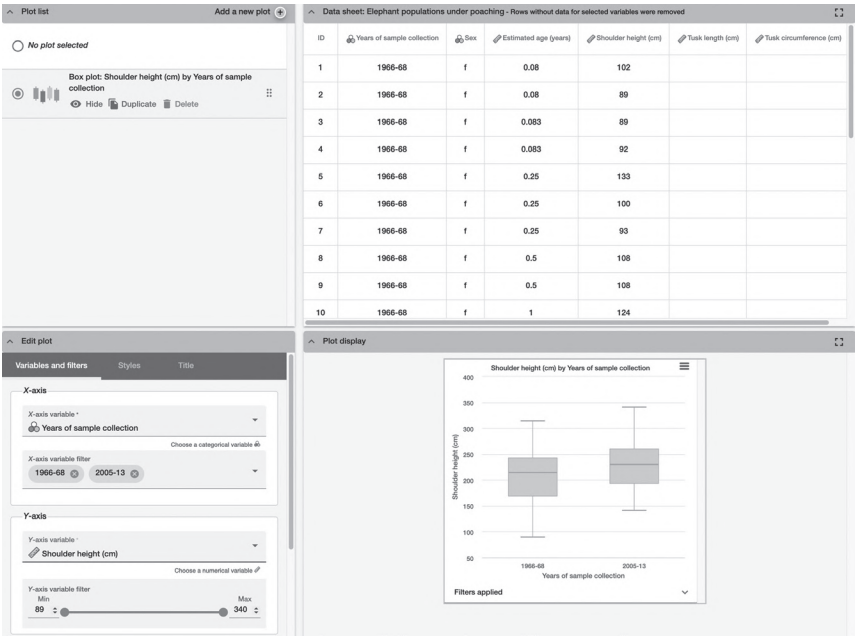

Figure 1. Example of the HHMI Data Explorer interface.

Data science has an important role to play in medicine and Nippon Medical School is underway in addressing this issue. The Center for Mathematics, Data Science, and Artificial Intelligence was established in part to introduce students to the connection between data science and medicine. The Department of Foreign Languages offers a scientific English writing course, which is unaffiliated with the center, but offers a place where students can put to practical use the data science skills they acquire. Encouraging collaboration between instructors of these respective courses could strengthen student outcomes and enhance motivation by creating a more cohesive learning experience. This collaboration would further encourage students to recognize the relevance of data science in medicine.

References

[1] Casà, C. *et al.* (2021) 'COVID-19 and digital competencies among young physicians: are we (really) ready for the new era? A national survey of the Italian Young Medical Doctors Association', *Annali dell' Istituto Superiore Di Sanita,* 57(1), pp. 1–6. Available at: https://doi.org/10.4415/ANN_21_01_01.

[2] Chow-White, P., Ha, D. and Laskin, J. (2017) 'Knowledge, attitudes, and values among physicians working with clinical genomics: a survey of medical oncologists', *Human Resources for Health*, 15(1), p. 42. Available at: https://doi.org/10.1186/s12960-017-0218-z.

[3] 'Data science' (2024) *Wikipedia*. Available at: https://en.wikipedia.org/w/index.php?title=Data_science&oldid=1260235404 (Accessed: 3 December 2024).

[4] Fröhlich, H. *et al.* (2018) 'From hype to reality: data science enabling personalized medicine', *BMC Medicine*, 16(1), p. 150. Available at: https://doi.org/10.1186/s12916-018-1122-7.

[5] Hulsen, T. *et al.* (2019) 'From Big Data to Precision Medicine', *Frontiers in Medicine*, 6. Available at: https://doi.org/10.3389/fmed.2019.00034.

[6] Iregbu, K. *et al.* (2022) 'Global health systems' data science approach for precision diagnosis of sepsis in early life', *The Lancet Infectious Diseases*, 22(5), pp. e143–e152. Available at: https://doi.org/10.1016/S1473-3099(21)00645-9.

[7] Khamisy-Farah, R. *et al.* (2021) 'Big Data for Biomedical Education with a Focus on the COVID-19 Era: An Integrative Review of the Literature', *International Journal of Environmental Research and Public Health*, 18(17), p. 8989. Available at: https://doi.org/10.3390/ijerph18178989.

[8] Siuly, S. and Zhang, Y. (2016) 'Medical Big Data: Neurological Diseases Diagnosis Through Medical Data Analysis', *Data Science and Engineering*, 1(2), pp. 54–64. Available at: https://doi.org/10.1007/s41019-016-0011-3.

[9] Subbiah, V. (2023) 'The next generation of evidence-based medicine', *Nature Medicine*, 29(1), pp. 49–58. Available at: https://doi.org/10.1038/s41591-022-02160-z.

[10] Subrahmanya, S.V.G. *et al.* (2022) 'The role of data science in healthcare advancements: applications, benefits, and future prospects', *Irish journal of medical science*, 191(4), pp. 1473–1483. Available at: https://doi.org/10.1007/s11845-021-02730-z.

（受付日　令和7年　2月　7日）
（受理日　令和7年　3月　7日）