Patterns

CellPress
OPEN ACCESS

People of Data

Navigating an early career in genomics and data science, written from the perspective of a current PhD student

Rebecca Tooze[1,*]
[1]Clinical Genetics Group, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK
*Correspondence: rebecca.tooze@rdm.ox.ac.uk
https://doi.org/10.1016/j.patter.2022.100548

Rebecca Tooze, a PhD student in the Oxford Clinical Genetics' group, discusses the importance and application for data science in her field. Using bioinformatic approaches, she analyzes whole-genome sequencing data from patients with craniosynostosis. In this paper, she comments on her current research and her opportunity as an editorial intern with *Patterns*.

What would you like to share about your background?

I am currently reading for a PhD in medical sciences at the University of Oxford, under the supervision of Professors Andrew Wilkie and Stephen Twigg, in the Clinical Genetics research group. In my short career, I have been fortunate to experience a range of fields including genetics, cancer biology, evolution, and microbiology; this diversity has afforded me the chance to learn a variety of molecular biology techniques, and my scientific interests have evolved in parallel toward projects with clinical application. My PhD project explores the genetic causes of craniosynostosis (the premature fusion of one or more of the cranial sutures of the skull). My research is a split of computational and wet-lab work, which allows me to continue developing my bioinformatics skills and molecular biology techniques while undertaking research with direct clinical benefit.

What motivated you to become a (data) researcher? Is there anyone/anything that helped guide you on your path?

Given the size of the human genome, data science methods allow us to accurately align human genomes to a reference, highlight variation in datasets, and annotate datasets with known variables for pathogenicity. It would be a painfully long task to do these steps manually and this does motivate me to search for better ways to use data science in my research. I was also fortunate to spend 8 weeks on a bioinformatics course, run by the MRC WIMM Centre for Computational Biology. This allowed me to immerse myself in the art of using Python, R and Linux to analyze and present real world data. The skills and knowledge I gained on this course certainly opened my eyes to the power and utility of bioinformatics within my own research and aided my transition from being a basic research scientist to a more confident data scientist.

What is the definition of data science in your opinion? What is a data scientist? Do you self-identify as one?

For me, data science is an umbrella term for how we use mathematics and statistics, computational power, and scientific thinking to extract value from data. It encompasses aspects of artificial intelligence (AI), whereby human behavior or intelligence can be simulated using technology, and machine learning (a subset of AI), which allows the machine to "learn" from past data. By this broad definition of data science, I would consider all elements of scientific research as data science, by nature of designing, conducting, and analyzing experiments and results. Although I do not use AI or machine learning in my research, I regularly use bioinformatics approaches in my analysis of sequencing data and variant prioritization, examples of which can be found in our recent paper, published in *Genetics in Medicine*.[1]

Which of the current trends in data science seem the most interesting to you? In your opinion, what are the most pressing questions for the data science community?

Personally, I am most excited about the potential of applying machine learning in the practice of medicine, particularly in clinical diagnostics, precision medicine, and health monitoring. Although I appreciate the ethical dilemmas of integrating this technology into current practice, promising research has shown the value of applying AI for pattern recognition in pre-processing and analyzing digital medical images and for classifying disease subtypes. With the advent of smart watches, this also allows the possibility to use features such as built-in heart monitors to measure individual health metrics, holding the potential to highlight deviations from the "norm" at much earlier stages. Having said this, I think there is a long way to go before this can be implemented, particularly in terms of navigating use of personal data associated with wearable technology.

What is the role of data science in the domain/field that you work in? What advancements do you expect in data science in this field over the next 2–3 years?

Genomic research generates a substantial amount of data, with estimates that between 2 and 40 exabytes of data[2] will be achieved over the next decade from sequencing alone. Our ability to sequence DNA far outpaced our ability to decipher the genetic code, and finding

ways to extract the value from these data is therefore a very vibrant topic. In its simplest form, geneticists employ various bioinformatics tools (methods and software tools) that allow us to collect and analyze biological data. On a clinical level, the interpretation of the identified variants (~3 million) is certainly one of the most time-consuming aspects of rare disease diagnosis, and AI therefore holds promise to simplify and speed up this process. Already, machine learning has enabled more advanced variant prioritization algorithms, combining predictive features (including effect on protein structure, amino acid substitution, and rarity) to generate a combined annotation-dependent depletion (CADD) score. Moving forward, we may expect to see AI-based methods take this a step further and predict features directly from raw sequences without prior data labeling. I would hope that our ability to use data science techniques, particularly AI and machine learning, may in future help to develop predictive tools for non-coding and synonymous variants too.

What do you enjoy about working with *Patterns*?

I have recently been offered an internship with Cell Press, working as an editorial intern with *Patterns* and *iScience*. Given the relatively new era of data science, and the scale at which this field is growing, it is an exciting time to be involved in *Patterns* and experience firsthand the new technologies and methods that are being published. As a comparatively new Cell Press journal, it is rewarding to be involved in increasing its outreach and identifying researchers in this field.

What paper(s) in *Patterns* particularly drew your attention and why?

Two articles that I found particularly interesting published in the January 2022 issue of *Patterns*:

1. Mohanty et al., 2022, "Machine Learning for predicting readmission risk among the frail: Explainable AI for healthcare."[3] The authors have identified a use for data science in aiding preventative care, by establishing risk factors associated with patient readmission. One of the main concerns with using AI or machine learning is the potential for bias in the model owing to incorrect assumptions. Lack of explanation within models for the contribution of each feature toward healthcare predictions greatly reduces confidence in the output. Here, the authors have successfully used Shapley additive explanation to explore reasons for model predictions at a global level, investigating predictions across the entire dataset, down to a local level, exploring individual patient observations. Employing these methods within data science should increase confidence and accuracy of predictions, allowing AI to be incorporated into healthcare.
2. Santos et al., 2022, "Machine learning and network medicine approaches for drug repositioning for COVID-19."[4] I wished to highlight this paper owing to its relevance of drug repurposing in the age of a global pandemic. It seems critical to be able to develop tools that are capable for screening current antivirals, in the context of virus-specific proteins, and identify those that may be effective, subsequently, greatly reducing drug development time. The authors have provided a free interactive online tool that explores the interplay between the SARS-CoV-2 host proteins and biological processes and identifies drugs that disrupt these networks.

REFERENCES

1. Hyder, Z., Calpena, E., Pei, Y., Tooze, R.S., Brittain, H., Twigg, S.R.F., Cilliers, D., Morton, J.E.V., McCann, E., Weber, A., et al. (2021). Evaluating the performance of a clinical genome sequencing program for diagnosis of rare genetic disease, seen through the lens of craniosynostosis. Genet. Med. *23*, 2360–2368. https://doi.org/10.1038/s41436-021-01297-5.
2. Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., Schatz, M.C., Sinha, S., and Robinson, G.E. (2015). Big Data: Astronomical or Genomical? PLoS Biol. *13*, e1002195. https://doi.org/10.1371/journal.pbio.1002195.
3. Mohanty, S.D., Lekan, D., McCoy, T.P., Jenkins, M., and Manda, P. (2022). Machine Learning for predicting readmission risk among the frail: Explainable AI for healthcare. Patterns *3*, 100395. https://doi.org/10.1016/j.patter.2021.100395.
4. Santos, S.d.S., Torres, M., Galeano, D., Sánchez, M.d.M., Cernuzzi, L., and Paccanaro, A. (2022). Machine learning and network medicine approaches for drug repositioning for COVID-19. Patterns *3*, 100396. https://doi.org/10.1016/j.patter.2021.100396.

Rebecca Tooze

About the author

Rebecca Tooze received her BSc in biological sciences (molecular and cellular biology) from the University of Exeter in 2019. As part of her degree, she studied abroad at Monash University, Melbourne, under the supervision of Dr. Michael McDonald. Following this, she undertook a short internship at the Beatson Institute, CRUK, before returning to Exeter to complete her degree. She then went on to study for a PhD in clinical genetics, where she continues to develop her research into identifying pathogenic mutations in patients with craniosynostosis and understanding the effect of these variants at the RNA and protein level.