Journal of Mass Spectrometry and Advances in the Clinical Lab 24 (2022) 41–42

Contents lists available at ScienceDirect

Journal of Mass Spectrometry and Advances in the Clinical Lab

journal homepage: www.sciencedirect.com/journal/journal-of-mass-spectrometry-and-advances-in-the-clinical-lab

Editorial

Clinical Pathology and the Data Science revolution

ARTICLE INFO

Keywords
Data Science
Clinical Pathology
Laboratory Medicine

While clinical laboratory medicine has always been replete with data, there has been relatively little effort applied from within our discipline to leverage it for clinical, operational, and financial insights. While the concept of "Data Science" as a scientific discipline is not new [1], in the past decade the development, maturation and democratization open-source Data Science tools has made it possible for any determined laboratorian to incorporate their use into clinical practice. Data Science is a multidisciplinary field incorporating aspects of computer science, mathematics, statistics, and predictive analytics for the purposes of extracting actionable insights from large datasets produced by any business or scientific sector. As it pertains to healthcare, Data Science can be leveraged for everything from diagnostic decision support to workforce analysis to the automation of repetitive data entry tasks.

Sometimes the nomenclature used in Data Science is confusing as there are a number of frequently arising and overlapping concepts: data analytics, big data, artificial intelligence (AI) and machine learning (ML). The overlap makes precise definitions challenging.

Data Science itself has been defined by Kelleher and Tierney as "a set of principles, problem definitions, algorithms, and processes for extracting nonobvious and useful patterns from large data sets" [2] while Donoho defines it as "The science of learning from data; it studies the methods involved in the analysis and processing of data and proposes technology to improve methods in an evidence-based manner" [3].

Data Analytics can be considered a necessary subset of Data Science directed at performing data cleansing, data merging, descriptive statistical analysis, and data visualization tasks for the purposes of drawing meaningful insights. Data Analytics is more often focused on business intelligence rather than scientific inferences per se.

AI is a term coined in the 1950's which has become very broad in its meaning. Fundamentally, AI is any computational process that creates the *appearance* of human intelligence. One can therefore leverage Data Science tools to build components of any AI system. ML is a subset of AI wherein the computational algorithm is able to learn from experience (that is through the provision of new example cases) without being explicitly programmed to do so [4]. Lee and Durant give a simplified beginner's tutorial of ML applied mass spectrometry data with code provided in both python and R to suit the reader [5]. Like AI, the development of any machine learner will involve a large component of Data Science.

Big Data more or less refers to data sets that are large, unwieldly, and can be characterized by the 3V's of volume, velocity, and variety [6,7]. Volume indicates the amount of data generated overwhelms traditional software tools and specific strategies (e.g. distributed computing) are required to perform the computational tasks [8]. Velocity indicates that the data is generated and accumulates rapidly. Variety, in the healthcare context, is easy to understand: numerical results, narrative notes and reports in the clinical chart, images from scanned external reports, radiology, anatomical pathology, and sequencing data from genomic studies. Clinical laboratory data fulfills this definition of Big Data with sources including the laboratory information systems, electronic health record, analytical instruments, and other ancillary systems. These data are frequently non-standardized and unstructured in their formats, requiring data cleansing/merging ("wrangling") prior to analysis.

The data-science tasks intrinsic to laboratory medicine have created a need for the clinical laboratorian to adopt new tools which include programming languages (R, Python, Julia), literate programming tools (Markdown), and web-app development tools (Shiny, Dash) and deployment strategies designed for reliability and long-term stability (e.g. use of cloud infrastructure, containers). These tools allow software development to fill gaps where no commercial solutions exist. Method validation, QC/QA, data automation, instrument interfacing, automated reports, and dashboard creation are typical targets for application development in clinical laboratory medicine. Examples of these can be found in Haymond's article on creating dashboards for business intelligence [9] and Geistanger's automated workflow for stability data [10].

Every area of the healthcare system is increasingly affected by Data Science and as more data is generated and stored, so also develops the

Abbreviations: AI, Artificial Intelligence; ML, Machine Learning; MS, Mass Spectormetry.

https://doi.org/10.1016/j.jmsacl.2022.03.001

Available online 17 March 2022

need to leverage it for clinical and operational insights. The scientific leadership of Mass Spectrometry and Advances in the Clinical Laboratory (MSACL.org) recognized the need for Data Science education nearly a decade ago and began promoting it as a discipline in clinical laboratory medicine through short courses, seminars, and now a special issue in JMSACL. Many of the clinical laboratory Data Science thought-leaders have either taken or contributed to MSACL short courses and have supported this special issue through articles or peer review. As an emerging area of research and operational interest, the gathering of the insights and experiences of the laboratory community seems essential to the development of the next generation of clinical laboratory scientists. The goal of this Data Science Special Issue is to draw on the laboratory community's knowledge to showcase the wide variety of Data Science applications deployed in laboratory medicine.

We have gathered articles that span a large swath of Data Science applications in the clinical laboratory from the aforementioned article by Haymond that demonstrates creating business intelligence dashboards [9], automated specimen stability statistics by Geistanger et al. [10], reproducible manuscripts in R Markdown [11], and workflows for continuous [12] and indirect reference intervals [13] to the applications to mass spectrometry including mass spectrometry imagining by Shedlock et al. [14] and Balluff et al. [15], MS quality metrics by Wilkes et al. [16] and Pablo et al. [17], and MS ML with Lee et al. [5].

We hope that this issue serves to inspire others to engage with Data Science in their workplaces.

Respectfully,
Dustin Bunch and Dan Holmes

References

[1] C. Hayashi What is data science? Fundamental concepts and a heuristic example. Vol. Tokyo: Springer Japan, 1998. pp. 40-51.
[2] J.D. Kelleher, B. Tierney, Data Science. Cambridge, Massachusetts: The MIT Press; 2018. xi, p. 264.
[3] D. Donoho, 50 years of data science, J. Comput. Graph. Stat. 26 (2017) 745–766, https://doi.org/10.1080/10618600.2017.1384734.
[4] H.H. Rashidi, N.K. Tran, E.V. Betts, L.P. Howell, R. Green, Artificial intelligence and machine learning in pathology: The present landscape of supervised methods, Acad. Pathol. 6 (2019), https://doi.org/10.1177/2374289519873088, 2374289519873088.
[5] E.S. Lee, T.J.S. Durant, Supervised machine learning in the mass spectrometry laboratory: A tutorial, J. Mass Spectrom. Adv. Clin. Lab. 23 (2022) 1–6, https://doi.org/10.1016/j.jmsacl.2021.12.001.
[6] S. Dash, S.K. Shakyawar, M. Sharma, S. Kaushik, Big data in healthcare: Management, analysis and future prospects, J. Big Data 6 (2019) 54, https://doi.org/10.1186/s40537-019-0217-0.
[7] N.V. Tolan, M.L. Parnas, L.M. Baudhuin, M.A. Cervinski, A.S. Chan, D.T. Holmes, G. Horowitz, E.W. Klee, R.B. Kumar, S.R. Master, "Big data" in laboratory medicine, Clin. Chem. 61 (2015) 1433–1440, https://doi.org/10.1373/clinchem.2015.248591.
[8] M. Zaharia, R.S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M.J. Franklin, A. Ghodsi, J. Gonzalez, S. Shenker, I. Stoica, Apache spark: A unified engine for big data processing, Commun. ACM 59 (2016) 56–65, https://doi.org/10.1145/2934664.
[9] S. Haymond, Create laboratory business intelligence dashboards for free using r: A tutorial using the flexdashboard package, J. Mass Spectrom. Adv. Clin. Lab 23 (2022) 39–43, https://doi.org/10.1016/j.jmsacl.2021.12.002.
[10] A. Geistanger, K. Braese, R. Laubender, Automated data analytics workflow for stability experiments based on regression analysis, J. Mass Spectrom. Adv. Clin. Lab 24 (2022) 5–14, https://doi.org/10.1016/j.jmsacl.2022.01.001.
[11] D.T. Holmes, M. Mobini, C.R. McCudden, Reproducible manuscript preparation with RMarkdown application to JMSACL and other Elsevier journals, J. Mass Spectrom. Adv. Clin. Lab 22 (2021) 8–16, https://doi.org/10.1016/j.jmsacl.2021.09.002.
[12] D.T. Holmes, J.G. van der Gugten, B. Jung, C.R. McCudden, Continuous reference intervals for pediatric testosterone, sex hormone binding globulin and free testosterone using quantile regression, J. Mass Spectrom. Adv. Clin. Lab 22 (2021) 64–70, https://doi.org/10.1016/j.jmsacl.2021.10.005.
[13] D.R. Bunch, Indirect reference intervals using an R pipeline, J. Mass Spectrom. Adv. Clin. Lab 24 (2022) 22–30, https://doi.org/10.1016/j.jmsacl.2022.02.004.
[14] C.J. Shedlock, K.A. Stumpo, Data parsing in mass spectrometry imaging using r studio and cardinal: A tutorial, J. Mass Spectrom. Adv. Clin. Lab 23 (2022) 58–70, https://doi.org/10.1016/j.jmsacl.2021.12.007.
[15] B. Balluff, R.M.A. Heeren, A.M. Race, An overview of image registration for aligning mass spectrometry imaging with clinically relevant imaging modalities, J. Mass Spectrom. Adv. Clin. Lab 23 (2022) 26–38, https://doi.org/10.1016/j.jmsacl.2021.12.006.
[16] E.H. Wilkes, M.J. Whitlock, E.L. Williams, A data-driven approach for the detection of internal standard outliers in targeted LC-MS/MS assays, J. Mass Spectrom. Adv. Clin. Lab 20 (2021) 42–47, https://doi.org/10.1016/j.jmsacl.2021.06.001.
[17] A. Pablo, A.N. Hoofnagle, P.C. Mathias, Listening to your mass spectrometer: An open-source toolkit to visualize mass spectrometer data, J. Mass Spectrom. Adv. Clin. Lab 23 (2022) 44–49, https://doi.org/10.1016/j.jmsacl.2021.12.003.

Dustin R. Bunch[a,b,*], Daniel T. Holmes[c,d]

[a] *Department of Pathology and Laboratory Medicine, Nationwide Children's Hospital, Columbus, OH, USA*

[b] *Department of Pathology, College of Medicine, The Ohio State University, Columbus, OH, USA*

[c] *St. Paul's Hospital, Department of Pathology and Laboratory Medicine, 1081 Burrard St., Vancouver, BC V6Z 1Y6, Canada*

[d] *University of British Columbia, Department of Pathology and Laboratory Medicine, 2211 Wesbrook Mall, Vancouver, BC V6T 1Z7, Canada*

[*] Corresponding author at: Nationwide Children's Hospital, 700 Children's Drive, Columbus, OH 43205, USA.
E-mail address: Dustin.bunch@nationwidechildrens.org (D.R. Bunch).