IJSPT
INTERNATIONAL JOURNAL OF SPORTS PHYSICAL THERAPY

Hewett TE, Olsen G, Atkinson M. The Use of Big Data to Improve Human Health: How Experience From other Industries Will Shape the Future. *IJSPT*. 2021;16(6):1590-1594. doi:10.26603/001c.29858

Invited Clinical Commentary

The Use of Big Data to Improve Human Health: How Experience From other Industries Will Shape the Future

Timothy E Hewett[1] [a], Greg Olsen[2], Mark Atkinson[2]

[1] Hewett Global Consulting; Rocky Mountain Consortium for Sports Research, [2] Sparta Science

Keywords: Health, Data Science, Machine Learning, Technology, Biometrics, Human Performance, Artificial Intelligence, Medicine, Occupational Health, Sport, Military, Statistics

https://doi.org/10.26603/001c.29858

International Journal of Sports Physical Therapy

Vol. 16, Issue 6, 2021

'Data science' represents a set of mathematical and software development related techniques that are applied across a wide range of problems and industries. Practitioners of data science in human health-related domains typically see a world that differs substantially from practitioners in other domains such as advertising, finance, e-commerce, manufacturing, or social networking. This commentary discusses what those differences are (Project vs Product Focus, Independent vs Integrated Efforts, Causality vs Prediction Driven, Statistical vs Machine Learning Centricity) why they exist, and the future convergence that we believe is on the horizon. The concepts discussed can provide a starting point in which health and human performance-focused stakeholders can begin to align well-established data science applications from other domains to further enable innovative health and performance solutions.

INTRODUCTION

'Data science' represents a set of mathematical and software development related techniques that are applied across a wide range of problems and industries. Practitioners of data science in human health-related domains typically see a world that differs substantially from practitioners in other domains such as advertising, finance, e-commerce, manufacturing, or social networking. Here we discuss what those differences are, why they exist, and the future convergence that we believe is on the horizon.

DATA SCIENCE CULTURE IN HEALTH DOMAINS VS OTHERS

The application of data science to problems related to human health has been shaped by a long history with clinical research studies used to determine the effectiveness and safety of medical treatments and medications. The structure of these studies is guided by regulatory bodies like the US Food & Drug Administration (FDA), and disciplines like biostatistics and epidemiology evolved to guide the appropriate application of data science to medical problem domains - research studies are central to both disciplines.

This long association of health-related data science with research studies focused on medical treatments has created a data science culture with distinguishing characteristics. Though other domains like finance, e-commerce, manufacturing, advertising or social networking are not monolithic, their data science cultures show distinct contrasts with health domains. Table 1 highlights some key differences that are discussed in subsequent sections.

FOCUS: PROJECT VS PRODUCT

PROJECT FOCUS

Data science efforts in health focus on clinical research studies. Typically, a study is conducted to answer a specific set of questions in a fixed time period. The medical approach comes with constraints such as availability of data, clearly specified outcomes, available resources and a justifiable benefit for the effort.

PRODUCT FOCUS

In other domains, data science is most often applied as an *ongoing* component of product/service development and delivery. Data science is applied incrementally and opportunistically to meet evolving needs of products, and commitments to data science can be long term and open-ended.

a **Corresponding Author:**
Timothy E. Hewett
Tim.Hewett@gmail.com
Hewett Global Consulting
1516 4th Avenue NW, Rochester, MN

Table 1. The Two Data Science Cultures and their Key Differences

Health Data Science Culture	*Culture in Other Domains E.g. Finance, eCommerce, Manufacturing*
Discrete Project **Focus**	Ongoing Product **Focus**
Independent Data Science **Efforts**	Integrated Data Science **Efforts**
Causality-**driven**	Prediction-**driven**
Statistical Data Model Centric **Methods**	Machine Learning (ML) Centric **Methods**

SIGNIFICANT EXAMPLES INCLUDE:

- Fraud detection
- Recommendation engines
- Search and SEO
- Language translation services
- Self-driving vehicles & other machine vision applications

EFFORTS: INDEPENDENT VS INTEGRATED DATA SCIENCE

INDEPENDENT DATA SCIENCE EFFORTS

In health domains, the people who conduct the analyses are typically independent of the people who use the results generated from the analyses. In some cases, this is a mandated order to ensure the objectivity of the analyses. In general, data science in health domains is conducted by dedicated specialists who provide results through studies to practitioners such as doctors and other medical professionals. Often data science is conducted by academics who are incentivized toward peer-reviewed publications rather than other forms of result dissemination or implementation. Data science work is concentrated in a relatively small group of specialists.

INTEGRATED DATA SCIENCE EFFORTS

Product-oriented domains may also have dedicated data science teams that do project-oriented work, but that work is a small fraction of the data science work that is integrated into product delivery teams. The scope of 'data science' in a product effort is expansive and includes roles such as instrumentation & data engineering, Machine Learning (ML) engineering, & ML operations. The 'modelling work' constitutes only a small portion of the data science related work in product development and delivery organization.

DRIVERS: CAUSALITY VS PREDICTION

CAUSALITY-DRIVEN

A primary goal in medical research is to find ways to positively change health outcomes for populations or individuals. Bio-statistical inquiry, therefore, has focused on understanding cause and effect relationships, for example:

- Is the treatment effective?
- Does a medication produce adverse side effects?
- Does an environmental condition lead to disease?

The quest for understanding causal relationships strongly shapes how data science is applied. Common practice is to use randomized controlled trials to test easily understood and actionable hypotheses relating cause to effect. Determination of causality is quite difficult. Sometimes, the establishment of a causal relationship is simply not feasible - even when the data supports predictive relationships. Many health-related problems are complex with multifaceted relationships between dependent and independent variables that are subject to change over time (emergent).

PREDICTION-DRIVEN

In other domains, applications of data science are less preoccupied with causality - prediction is the primary focus. Consider relevant and important questions such as:

- Given a set of measured conditions, what is the probability of rain?
- Does a given transaction history imply potential fraud?
- What is the best match for a search term?
- Is the current state of the system anomalous?
- What does this photo most resemble?
- What additional product might this person add to their checkout cart?

Removal of a strict causality requirement opens the door to an expansive set of data science analysis options. Analysis can be entirely data-driven, i.e., the investigators can explore a dataset looking for patterns with no constraints on what can be a feature/factor/variable and learn to exploit those patterns as we see fit. Any method or approach that shows predictive power is a viable option for data science pursuits.

MODELING APPROACHES: STATISTICAL DATA VS MACHINE LEARNING

STATISTICAL DATA CENTRIC MODELS

Statistical data model-based approaches are entrenched in health-related data science culture. These approaches are characterized by the following:

- Preference for a set of familiar linear algorithm-based methods that support a direct approach to explanation.
- Exclusion of 'black box' prediction techniques.
- Focus on a small number of predictors that are 'well understood.'
- Data set sizes are small to medium.

- Statistical significance criteria utilized.

MACHINE LEARNING CENTRIC MODELS

The machine learning culture present in other industries is very diverse but several aspects contrast with statistical data model cultures, which include:

- A preference for non-linear methods, particularly ANNs (Artificial Neural Networks),
- Separation of prediction and explanation into independent analyses.
- Use of many features and use of machine learning-based feature learning.
- Many applications have access to 'big data.'

In healthcare, it is generally assumed that a data science model needs to be understandable and trustworthy by people (like doctors) who consume those models but are not able to perform the data science analysis themselves. This constraint makes the use of new mathematical techniques or more complex approaches to feature learning or engineering quite difficult. Many product-oriented domains are free from this constraint and the driving goal is simply maximum predictive power. Also, many applications in non-health domains are not governed by the oversight and regulation present in health domains whereby the ability to explain results is not explicitly required in all cases.

Some differences in culture are simply due to an earlier and less constrained move into new machine learning techniques in domains outside of healthcare. Healthcare draws more strongly from statistics, biostatistics, and epidemiology academic programs while other domains draw more from computer science and engineering; the latter were earlier to introduce machine learning into curriculums and less informed by prior statistical methodological practices. Because of the long history, healthcare data science practice has evolved more slowly and cautiously than many other domains. 20 years ago, Leo Breiman highlighted this reluctance to diversify methodologically in his famous paper about the "two cultures" of statistical modelling,[1] which is still a subject of debate today.

One other driver for the use of machine learning are 'big data' situations where voluminous streams of data are available for model learning. Whether it be sensor or transaction data, large volumes support accurate pattern recognition, and an ongoing stream can support continuous model improvement. Much of the rapid progress in machine learning applications is directly attributable to 'big data', and more recently this type of data availability is becoming increasingly relevant in healthcare domains as well.

OPPORTUNITY THROUGH CONVERGENCE

Existing healthcare related data science culture exists for a reason. The unique demands in ensuring safe and effective medical treatments make clinical research studies a necessity, and warrant caution with the introduction of new techniques and technologies. However, fuelled by the exponential expansion of health-related data and of new uses of data, practices from other domains are inevitably making their way into a broader definition of health data science.

There are already many examples where the characteristics discussed previously (product centricity, integrated data science, and ML-tech centricity) are already present in health-related applications, for example:

- **Wearables and associated health applications**: Wearable devices produce big data streams that support continuous model learning and a host of application possibilities for sport performance, sleep monitoring, health assessment, e.g.
  - Oura, Apple Watch, Fitbit, Garmin, Amazon Halo
- **Automated assistance**: Leveraging advanced NLP (Natural Language Processing) capabilities, bot-based applications are emerging to provide assistance to people in many contexts including mental health.
- **Smart exercise equipment and applications**: Highly instrumented training systems that use ML/AI (Machine Learning/Artificial Intelligence) to guide the user toward better health outcomes and improved physical performance, e.g.
  - Peloton, Mirror, Blast
- **Radiology & pathology**: Radiologists and pathologists can now leverage a rich set of image recognition technologies and approaches from ML/AI that were initially developed on other problems like NN-based image recognition and shape detection in consumer and manufacturing applications.
- **Drug discovery**: While clinical trials are still required to approve drugs for use, the discovery of potential drug targets is being done using ML/AI techniques initially developed for other purposes such as search or data mining in other domains.

Applications of data science such as these represent only a first phase of evolution where health data science retains some distinct characteristics, but in aggregate looks more and more like other application domains. The healthcare community requires new big data approaches to move forward into a more efficacious and cost-effective future. With Artificial Intelligence approaches, the "sky appears to be the limit,' but we must learn from the experiences and advances of other industries to make the most of these new technologies safely and with optimal health and medical outcomes.

DISCLOSURES

Two of the authors, Olsen and Atkinson, are employees of Sparta Science and Hewett is Chairman of the Scientific Advisory Board of a data science company.

Submitted: October 28, 2021 CST, Accepted: November 07, 2021 CST

REFERENCES

1. Breiman L. Statistical Modeling: The Two Cultures.
Statist Sci. 2001;16(3):199-215,. doi:10.1214/ss/1009213726