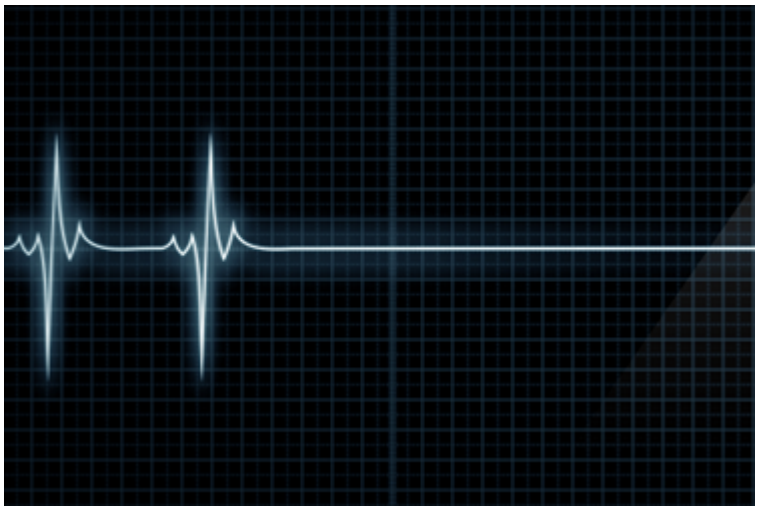
BIOPROCESS ONLINE

Guest Column | June 18, 2019

AI, Data Integrity, & The Life Sciences: Let's Not Wait Until Someone Dies

By Kip Wolf, Tunnell Consulting, @KipWolf



The idea for machines that can think became the topic of science fiction in the early parts of the 20th century and made for interesting reads. Science caught up and the term "artificial intelligence" (AI) was coined by John McCarthy at the Dartmouth Summer Research Project on Artificial Intelligence (DSRPAI) in 1956, where the first AI program, the *Logic Theorist*, was presented by Allen Newell, Cliff Shaw, and Herbert Simon.[1]

AI research flourished in the early years until it was slowed by limits in computational power, but it was reinvigorated in the 1980s by both computational tools and investment when John Hopfield and David Rumelhart popularized deep learning techniques that allowed computers to learn using experience.[1] The next limitation to AI advancement was in computer storage, which by the late 1990s was no longer a problem, as storage advancements produced cheap and ubiquitous solutions. In our modern world, we carry devices in our daily lives that dwarf the storage capability of supercomputers of only a few decades ago. AI has now gone mainstream, leaving the labs and coming into our living rooms with intelligent assistants (i.e., Alexa and Siri) and smart TVs. AI is on the news and on our tongues, as scarcely a week goes by without a television commercial or someone in our social circles mentioning AI. But what is AI and how might it impact our lives when applied to life sciences?

Garbage In, Garbage Out

For the sake of this discussion, we can agree that AI may result in predictions, classifications, and decisions from computational analysis of large data sets that is based on machine learning (ML) from representative data sources and further informed by said data and related results. In this context, AI may, for example, present significant potential for improving efficiency of research and development activities such as discerning viable drug targets for further investigation. Or, AI may offer greater capacity for manufacturing by reducing the potential for defects and accelerating product review, release, and disposition for shipping through the supply chain. However, there remains great risk with this potential for great reward, as mistakes or losses caused by poor AI results could cause negative impacts on public health.

One AI challenge today is in developing and managing ML and AI to handle the great volume of disparate and non-standardized data that is available. AI is being rapidly adopted in our lives, made obvious in examples of consumer electronics. Karan Bedi, COO of Blaupunkt Televisions India, reports that "consumer goods companies are leaving no stone unturned to empower their products with digital and AI technologies" and "many household appliance manufacturers integrate the Internet of Things (IoT) and AI in the household products."[2] A common example is the smart TV, the global unit share of which rose to over 70 percent of TVs sold in 2018, up from 55 percent in 2015.[3]

Consider how bad data could affect the AI experience. Data integrity and data quality play key roles in AI results. Poor quality input may produce unexpected or erroneous AI output. Take, for example, the use of a smart TV where Netflix data was carelessly entered (e.g., randomly selected programs of interest) or a Hulu account login was entered by a houseguest. When an algorithm used for targeted advertising or suggested programming in either of these services is applied to the data set, the results might have no relevance whatsoever to the current viewer. While this may be irritating or unhelpful, it is not life-threatening.

However, careless data entry or incorrect data sets related to life sciences applications could have consequences that include mortality. "Machine learning algorithms are very dependent on accurate, clean, and well-labeled training data to learn from so that they can produce accurate results," says Ron Schmelzer.[4] During ML, biased or erroneous inputs can cause inaccurate or anomalous outputs that have no relevance to the patient at hand. While mistakes are unlikely, the acceptability of error drops dramatically when considering any negative impact to patient health and public safety. Viewing Netflix programming advertisements that are of no interest is one thing – being dosed with the incorrect medicine is something else entirely.

Begin With The End In Mind

A market has emerged for data preparation solutions (including ClearStory Data, Datameer, Datawatch, Melissa Data, Oracle, Paxata, SAP, SAS, TIBCO Software, Trifacta, and Unifi Software) that perform data wrangling, data cleaning, and data preparation to enable ML and AI. In fact, "the vast majority of machine learning project time" is taken up by these activities.[4] However, at the rate that data is being created, it is likely that the ability to prepare data will be outstripped by the backlog of data to be prepared.

Data preparation continues to require some level of human interaction. At a minimum, a human must configure the specification for data transformation during ETL (extract, transform, and loading) when gathering data from multiple sources or migrating data to a central data store. Yet data wrangling represents potentially greater human involvement, as context may be necessary to perform advanced processing for transformation of robust data.

Whether data preparation solutions can keep up or not, the argument remains for improving the integrity and quality of data as it is created, rather than attempting to clean it up later as it is being prepared for ML and AI. This is accomplished in large part through data management and information governance where data integrity and data quality are central tenets. It is incumbent upon those creating AI for life sciences solutions to apply the greatest attention to data integrity and data quality to mitigate the risk of negative impact on patient health and public safety. AI for life sciences solutions are held to a higher standard than in other industry sectors.

Data Integrity A Critical Success Factor

Data integrity and data quality are critical success factors for AI solutions in life sciences. Standards for and verification of data integrity and data quality must be elevated for data sets where ML/AI is intended to be applied. Simply performing computer system validation (CSV) or managing computer systems under CGMP conditions is not enough to ensure data integrity and data quality.

Data integrity and data quality must be common themes in a mature quality management system and proactively integrated into data management and information governance as a core business activity. We find that when firms understand that data integrity and data quality are critical success factors, the result is a competitive advantage. Fewer human errors may occur, and investigations may be completed more rapidly and successfully when they do. M&A activities are made more effective and efficient as due diligence is more easily facilitated and valuation is more clear with defensible data and with human resources who understand and can explain it. These success factors lead to better AI results with improved ability to provide products to patients and increased value to owners and shareholders.

Now that AI has become more common, we are presented very directly with practical and ethical questions as we "allow AI to steadily improve and run amok in society."[1] When will an AI result based on bad data end in a consequence of human injury or even death? At what point will malicious intent be involved to influence those AI outcomes to result in injury or death? Remember the "Tylenol murders of 1981" and how the resulting regulatory and industry actions changed forever how we package medicines.[5] Will we behave reactively, waiting for a "Tylenol-level event" to force us to govern ourselves and the bad data we are pumping through AI? Or will we behave proactively to fervently manage our data to prevent negative impacts on patient health and human life?

References:

1. Anyoha, R. (2017, August 28). The History of Artificial Intelligence. Retrieved June 8, 2019, from Science in the News website: http://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/
2. Bedi, K. (2019, February 21). How Artificial Intelligence is Reinventing Consumer Electronics Segment. Retrieved June 7, 2019, from Entrepreneur website: https://www.entrepreneur.com/article/328400
3. Global smart TV market share 2015-2018. (2018, July). Retrieved June 7, 2019, from Statistica website: https://www.statista.com/statistics/889000/worldwide-smart-tv-market-share/
4. Schmelzer, R. (2019, March 7). The Achilles' Heel Of AI [Forbes]. Retrieved June 7, 2019, from https://www.forbes.com/sites/cognitiveworld/2019/03/07/the-achilles-heel-of-ai/#d08829c7be7e
5. Markel, H. (2014, September 29). How the Tylenol murders of 1982 changed the way we consume medication. Retrieved June 8, 2019, from PBS NewsHour website: https://www.pbs.org/newshour/health/tylenol-murders-1982

About The Author:

Kip Wolf is a principal at Tunnell Consulting, where he leads the data integrity practice. Wolf has more than 25 years of experience as a management consultant, during which he has also temporarily held various leadership positions at some of the world's top life sciences companies. Wolf temporarily worked inside Wyeth pre-Pfizer merger and inside Merck post-Schering merger. In both cases he led business process management (BPM) groups — in Wyeth's manufacturing division and in Merck's R&D division. At Tunnell, he uses his product development program management experience to improve the probability of successful regulatory filing and product launch. He also consults, teaches, speaks, and publishes on topics of data integrity and quality systems. Wolf can be reached at Kip.Wolf@tunnellconsulting.com.