Data Quality for AI in Healthcare

FEBRUARY 2021

XAVIER HEALTH

TABLE OF CONTENTS

ABOUT XAVIER HEALTH

Xavier Health is more than an organization. It is a community of hundreds of FDA, industry experts, thought leaders and academics. Xavier Health was formed in 2008 as an outreach of Xavier University charged with making a difference in the pharmaceutical, medical device, and combination products industries. Our mission is inspiring collaboration, leading innovation, and making a difference in all that we do.

XAVIER HEALTH

www.xavierhealth.org

ACKNOWLEDGMENTS

This paper was developed under the leadership of the Xavier Health program at Xavier University in partnership with industry professionals, as a planned output from 2019 Xavier AI Summit. We would like to thank everyone who contributed to the creation and the review of this paper—without their work, this paper would not have been possible[1]. We also want to acknowledge the significant contributions from the following people:

- Bob Banta, Eli Lilly
- Gilbert Cortes, J&J
- Steve Frigon, The Christ Hospital
- Lacey Harbour, Lima Corp
- Tripti Kataria, MD, MPH, FASA
- Eileen Koski, IBM
- Betsy Macht, Walden University
- Mac McKeen, Boston Scientific
- Sundar Selvatharasu, Sierra Labs
- Sunnie Southern, Viable Synergy
- Scott Thiel, Guidehouse
- Bradley Merrill Thompson, Epstein Becker & Green
- Paul Westfall, AMA
- May Yamada-Lifton, SAS
- Zina Manji, GSK
- Geoff Waby, Advanced Testing Laboratory

Our hope is that this paper provides the foundation for new learnings and best practices in this rapidly evolving field to help deliver the promise and potential of AI.

Pat Baird, Philips

Rohit Nayak, ERS Inc.

Jackie Karceski, CAI

[1] Please note that the opinions and viewpoints expressed by the contributors do not necessarily reflect the opinions and viewpoints of their organizations.

1. INTRODUCTION

Background

Over the past few years, Artificial Intelligence (AI), and more specifically Machine Learning (ML) technology has experienced rapid adoption in the healthcare space as tools for diagnosis and decision-making. Such tools are intended to address challenges in the health care system to both process and the application of rapidly proliferating medical findings to practice, as well as to deliver on the promise of personalized and precision medicine.

The classic computer science idiom of "garbage in = garbage out" certainly holds true for ML systems; the quality of the data used to develop and test the system has a significant impact on the quality of the system output. There are many examples in the press where poor data quality led to poor recommendations and even led to instances of discrimination against certain patient populations, due to bias in the data that was used to develop the system. This paper attempts to describe and address these potential data quality issues.

Overall Goals of this Whitepaper

This paper is intended to open a discussion of current approaches for Data Quality by highlighting commonly used systems-development constructs oriented towards healthcare. By highlighting a series of considerations and questions, the reader is enabled to make their own conclusions.

This document addresses data quality through the entire data lifecycle, including characterizing the data to be acquired, potential bias in the data, data selection, cleaning and manipulation, and storage. Additional topics such as data streaming, risk management, etc., are also discussed.

This paper assumes that the reader is familiar with the basic principles of quality management systems and software development processes, and therefore it will not discuss such topics as the importance of having an approved plan, the need to document decisions, the need for change control, etc.

This paper is not intended to be a standard, nor is this paper trying to advocate for one and only one approach to ensuring data quality. This paper is also not intended to evaluate existing or developing regulatory, legal, ethical, or social consequences of these approaches.

Audience and Stakeholders

The intended audience of this paper is broad, and includes developers, implementers, researchers, quality assurance, regulatory affairs, validation personnel, business managers, regulators and end-users faced with the challenge of assessing the quality of and building trust in AI healthcare applications.

2. DATA CHARACTERIZATION

KEY POINTS TO HIGHLIGHT IN THIS SECTION

To get the benefits of ML, the data examined needs to embody key characteristics. A lack of any of these characteristics could adversely affect the output and outcomes determined from the data. The characteristics of data are outlined as "the V's": volume, velocity, viability, variety, veracity, volatility. The quality and quantity of data is critical to enabling the application to learn and reapply.

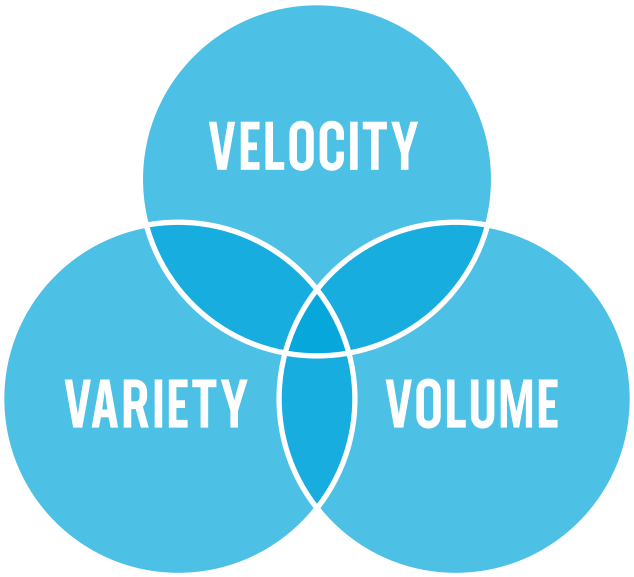
Discussion

There are many characteristics that can affect the quality and usefulness of data. The data values might not correctly describe the patient condition, or the data values might be no longer relevant, or the data itself might be fine but there might not be a large enough sample size to have confidence in the results, etc. Sometimes these characteristics are called the "3 V's of Big Data"[2]—even though this advice has evolved over the years to include additional "V's."

The V Factors

The purpose of this section is to provide a list of characteristics that should be considered when evaluating data quality. Note that not every characteristic applies to every application.

1. **Volume:** How much data is there? If you have too little data, the application will not be able to robustly learn.
2. **Velocity:** How quickly is the data being created? If you have too little data, how long will it be before you have an adequate amount of data? Is the data no longer relevant?
3. **Velocity Δ:** Is the velocity of data creation accelerating or decelerating? Are the changes foreseeable?



4. **Variety:** How many data sources are there? Does this application rely on only one source of information? Is the data adequately representative of the intended use or user group? If there are different data sources, what are the unique characteristics of those sources. For example, although patient data may be pulled from multiple hospital locations, are those hospitals part of the same network?
5. **Veracity:** Why do you think you can trust the data? It is not uncommon to spend a significant amount of time cleaning up data before it is used.

[2] "Volume, Velocity, Variety: What You Need to Know About Big Data" https://www.forbes.com/sites/oreillymedia/2012/01/19/volume-velocity-variety-what-you-need-to-know-about-big-data/?sh=6a0609b1b6d2

6. **Validity:** Are the data values correct? Are they timely? Describe the protocol used to collect this data, explaining 'When', 'How' and 'by Whom' data was gathered.
7. **Viability:** How is the data relevant to the use case?
8. **Volatility:** How often does the data change? Describe how long is it relevant, how long to store the data before archiving or deleting.

Some additional aspects of Data Characterization are discussed in the Data Bias & Selection section (Section 3), including:

1. What demographic does this serve?
2. Does the patient/study population match the intended use and indications for use?
3. Does it use any 're-calibration' or 'compensation' technique(s) to adapt input data to that of a target demographic setting?

Guidelines

The following are some recommended guidelines to leverage the V's:

1. An analysis should be performed regarding the impact of the relevant data characteristics on product performance. This should include the amount of training data available (Volume) and how that impacts confidence in performance. This would include an analysis regarding the applicability (Viability) of the data (e.g., is the data no longer relevant?), the number and Variety of data sources used and how that can impact performance, the Veracity of the data and data cleaning steps that need to be performed, if applicable, ensuring the Velocity of the incoming data is sufficient for the application needs, factors impacting the Validity of the data, and how often the data changes and what the impact is of that change (Volatility).
2. An analysis should be performed of the intended patient population vs. the development/test population, and differences should be discussed and analyzed for their impact, as well as any process steps used to compensate for changes.

FDA'S ACTION PLAN INCLUDES WORK WITH XAVIER WORKING TEAMS

FDA issued the "Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan" from the Center for Devices and Radiological Health's Digital Health Center of Excellence.

As part of this Action Plan, FDA is committing to deepening its work in collaborative communities, which includes the Xavier Working Teams, in order to encourage consensus outcomes that will be most useful for the development and oversight of AI/ML-based technologies. →Get engaged in moving the industry

3. DATA BIAS & SELECTION

KEY POINTS TO HIGHLIGHT IN THIS SECTION

Bias is inherent in data gathering, management and interpretation. Bias must be understood, mitigated, and accounted for. Bias can be related to the data itself, the interaction with the hardware, or the user. The demographics of the people involved at all stages can introduce unintended and intended bias.

Definition of Bias

There are different types of bias that can be introduced to any project. Every individual on the project team can hold bias that may potentially influence decisions. The definition of the intended population can have inadvertent bias, and the data can potentially reflect that bias. In this section our focus is on the potential bias introduced to the data associated with data-driven artificial intelligence (AI) applications.

Data Bias

Data bias can be introduced during the data gathering process. Over the lifecycle of product development bias could be introduced by product changes, software complexity, human resources, and inadequate management of data and programming risks[3]. There is the potential for inadvertent incorporation of unconscious societal bias into the data set.

The demographics associated with the source of the data set can potentially introduce bias. In defining the AI application consider which demographic the product serves, whether there is a population match in alignment with the intended application(s), and whether re-calibration or compensation technique(s) were used to adapt input data to align correctly with the target demographic setting.

Components of Data Bias

In evaluating and managing bias, the factors to consider include understanding, mitigating, and accounting for bias[4]. In this section, the components of data bias are considered. Data can be original, collected specifically for the development goal. Data can also be sourced through third-party entities, however, the ability to accurately assess the following points may be limited with third-party data sources.

The following should be considered when assessing an application for bias:

1. What kinds of bias might exist in the data? Have you considered data generation and collection parameters that could add bias?
2. What have you done to evaluate the data for bias, and how does it affect your model?
3. How have you processed the data to mitigate the influence of bias?
4. What are possible risks due to bias and what is your mitigation plan?

[3]Malika, V. & Singh, S. (2018). Quality Assurance & Risk Management in Computing Environments ICCM. 222-226. https://www.SSRN.com/link/ijisms-pip.html

[4]Ntoutsi, E. et al. (2020, February 3). Bias in data-driven artificial intelligence systems—An introductory survey. WIREs Data Mining and Knowledge Discovery, 10: e, 1-14. https://doi.org/10.1002/widm.1356

5. What residual bias might remain and how should users take this into account? How do you disclose the residual bias to the end users?

6. What factors affect data quality and what has been put in place to control that quality?

7. Has the data been evaluated for data completeness? (in terms of ranges, variations, outliers, missing values, etc.)? Has the data been reviewed to remove duplicate or questionable entries?

8. How will different hardware platforms affect the quality of the development, test, and use data, and how has this been accounted for? (For example, different model cell phones have different resolution cameras). What kind of normalization technique is used in data capture from different hardware specifications?

9. What is the method of ground truth determination? What type of data classification and data labelling processes were followed? (e.g. (i) manual, (ii) semi-automated or (iii) fully automated). Define the existing 'gold/reference/benchmark standard' being followed.

10. How do you plan to maintain the quality of the data relative to potentially changing use cases or users (e.g., changes in medical practice may result in the current model no longer being valid)?

Data Selection Bias

In the development process inadvertent bias could be introduced by the project team members due to a lack of awareness of the issues of other groups and stakeholders that may not be directly represented in the demographics of the project team or the data (e.g., race, gender, comorbidities, etc.). Both selection bias and exclusion bias could apply within the definition of the data set. Different societal groups could be over or underrepresented in the data used for development. This inadvertent bias could introduce risk to the user or service provider from bad (i.e. incomplete) data. The bias introduced through lack of awareness of the issues of different user groups could also manifest as training bias.

The following are some considerations that a project team can leverage to address data selection bias:

- Where was the data collected?
- When was the data collected (temporal disassociation)?
- Has inadvertent gender bias been created?[5]
- Is there the potential to choose data from a tertiary care center that may be skewed?
- What is the testing algorithm?
- What parameters were set for collecting?
- How was the data managed during the preprocessing, in processing, and post processing phases to mitigate bias?

[5] Sun, T. et al. (2019, January). Mitigating gender bias in natural language processing: Literature review. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Florence, Italy), P19-1159, 1630–1640. Association for Computational Linguistics. doi: 10.18653/v1/P19-1159

Another consideration in assessing the risk for data bias is in the design of the device and the interaction between the device and the user. For example, racism and sexism in medical research has been studied and documented showing that historically the study population for medical studies has included white males introducing inaccuracies and bias in the application and treatment of women and minorities[6][7][8]. Another example includes ensuring that the design of medical equipment includes the engineering and ergonomic analyses to meet user requirements, which is important to the overall functionality of the device[9]. If the user interface does not enable accurate collection of data, the data collected could contain bias impacting the learning algorithm interpretation of the data and the analysis of data by researchers.

JOIN THE CONVERSATION

Follow the Xavier AI LinkedIn group, Artificial Intelligence - Advancing the Pharmaceutical and Medical Device Industries, at **https://www.linkedin.com/groups/12078247/**

XAVIER HEALTH

[6] Criado-Perez, C. (2019, June 15). End medical gender bias. New Scientist, 242(3234). https://doi.org/10.1016/S0262-4079(19)31077-2

[7] Dennis, G. C. (2001, March). Racism in medicine: planning for the future. Journal of the National Medical Association, 93(3 Supplemental), 1S-5S. https://pubmed.ncbi.nlm.nih.gov/12653392/

[8] Hamberg, K. (2008). Gender bias in medicine. Women's Health, 4(3), 237-243. https://doi.org/10.2217/17455057.4.3.237

[9] Martin, J.L., Norris B.J., Murphy E., Crowe, J.A. (2008) Medical Device Development: The challenge for ergonomics. Applied Ergonomics 39 (3), 271-283. https://doi.org/10.1016/j.apergo.2007.10.002

4. DATA CLEANING AND MANIPULATION

KEY POINTS TO HIGHLIGHT IN THIS SECTION

Clean data is the foundation of safe and effective medical applications. Data cleaning is often the most time-consuming activity in preparing an application. Data cleaning must preserve integrity of the information contained in the data set. Data cleaning must be documented in an audit trail and validated. Real-world patient variability drives data set variability and must be considered to avoid inappropriate deletion of "outliers." A Data Quality plan and Data Manipulation plan are required to preserve data integrity.

Discussion

Data sources are natively problematic. They are often incomplete, include duplicates, lack consistency and include data entered erroneously. The data might include multiple data types, or data collected with different units of measure. In other cases, different devices may have generated the data making comparisons difficult (e.g., devices having different precision and accuracy capabilities.) The data may simply be corrupt in other ways (e.g. mismatch of patient identifier and result.)

Data cleaning is a remedial step performed once the desired data characterization is understood. Data cleaning addresses the objective errors and inaccuracies that we are able to detect in the data, as well as at least some of the biases that can be structurally corrected, for example, by separation that allows for later independent analysis.

After the data has been flagged as clean, the data sets need to be manipulated to put them into a form that can be analyzed. Data cleaning and manipulation, while they can be very tedious, often constitute most of the work performed to or with the data.

Factors to Consider

There are several objectives to the data cleaning and manipulation steps.

First, the goal of the cleaning step is to identify and correct any errors in the data relative to the intended use of the data. The identification step often involves visualizing the data, manually reviewing samples of the data, and using statistical techniques to find the gaps and other errors in the data set. Once errors are identified, the cleaning steps must be effective in remediating the errors to set up subsequent analysis. This analysis needs to be documented, reviewed and approved to ensure the analyst is correct in the assumptions and decisions being made, and any residual impact is understood.

Second, these steps must preserve the integrity of the information and its suitability for the intended use of the model. It is quite easy to, when correcting for one type of error, introduce a new error. Newly introduced problems may lead to errors in the output of the model. This also includes appropriate change control and storage of the data to maintain data integrity and accessibility. There may also be a need to consider confidentiality requirements associated to the data (e.g., personally identifiable information).

Third, the manipulation steps must put the data in a suitable format for effective analysis according to a prespecified plan.

Fourth, the developer must have a plan for validating the cleaning and manipulation steps to ensure that they accomplish their purposes.

Fifth, the developer should consider the benefits to downstream users of attaching a "fingerprint" to the data set. This may include information such as the data source demographics, data ownership, data purpose and limitations. Subsequent users who modify the data may add key details to the fingerprint. This information may help considerably when future users need to interrogate the data during problem solving activities.

Finally, these cleaning and manipulation steps must assure that the analysis is transparent and documented for anyone charged with auditing the work. Quality assurance, and good business practices, requires that work be done in a way that can be checked by others.

Guidelines

The intended uses, data sources and types of data used in machine learning for medical applications are quite varied and no single approach will always work. Thus, the following are simply general best practices to be observed.

1. The developer should write the intended use of the model in sufficient detail prior to the cleaning and manipulation steps so that the statement can form the basis for deciding the suitability of cleaning and manipulation steps.

2. The developer should thoroughly explore the data visually, manually and through statistical tests to identify problems with the data and anomalies that need to be addressed in the data cleaning steps. A thorough understanding of the data is required before cleaning and manipulation can be done.

3. The developer should thoroughly document the specific data problems found and the solutions to those problems used, with each step explained and justified for the intended use to allow for effective auditing. For example:

    a. Missing data: The developer should document what types of, and to what extent, the data are missing, the method chosen for addressing the missing data, and a justification for that method with reference to the intended use.

    b. Outliers: The developer should not discard outliers or other data without clearly documenting the investigation and rationale for the deletion. Personalized medicine is now teaching us that patients are much more individualized than we first understood. When we look at large data sets, we may see lots of variability patient to patient. The temptation is to characterize some of those as outliers and simply delete them. The problem is that the cleaned output might simply be wrong. It might mask the highly idiosyncratic relationship between treatments and individual patients.

4. Data Quality Plan: Going forward, the developer should generate a Data Quality Plan.

    a. Based on the intended use, the developer should set key metrics for the quality of the data that need to be achieved.

    b. The plan should include standardized procedures to manage the importation of new data to make it easier to catch duplicates and corrupt data before they are merged with other data.

    c. The developer should standardize the process for all future cleaning so that it is performed consistently.

    d. This plan should specify how the developer will monitor the quality of the data over time. The developer should build test steps to flag future data deviations that are not within expected ranges.

5. Data Manipulation Plan: The developer should develop a Data Manipulation Plan. The data manipulation steps in the plan should flow from the intended use, and lead to a final presentation of the data that allows for the appropriate analysis.
6. The developer should convert the manipulation steps to durable and reliable production code.
7. Validation: The developer should validate the effectiveness of the final data cleaning and manipulation steps, both qualitatively and quantitatively.
    a. Qualitative validation typically includes visualizations meant to identify the presence of key characteristics of the data.
    b. Quantitative validation includes key statistical tests to ensure that the ultimate data are in a form appropriate for the next step in the model development process.
    c. The developer should specifically address the risk that bias may be introduced at this cleaning and manipulation phase and should explain how either no bias is introduced or any bias that might be introduced is appropriately mitigated.
8. Someone outside the development team should audit the Data Quality Plan, the Data Manipulation Plan, and the results of the validation steps.
9. The developer should periodically repeat the validation steps.
10. The periodic validation reports, any changes to the Data Quality Plan and the Data Manipulation Plan should be periodically re-audited.

XAVIER AI WORKING TEAMS

Through the Xavier Artificial Intelligence Initiative, chartered working teams have been established during the face-to-face setting of the annual Xavier AI Summit. The teams that are currently formed are as follows, and are accepting new members:

AI in Operations (AIO) Team
The AIO Team is an organized, cross-industry discussion group of FDA officials and industry professionals working to increase the predictive assurance of product quality across all operations through the power of AI. →Learn more

Good Machine Learning Practices (GMLP) Team
The GMLP Team is bringing the world of AI activity into one place in order to increase the awareness of good work that has already been done and to collaboratively further solutions that address challenges related to AI implementation across the industry. →Learn more

AI at the Point of Care (AI@POC) Team
The AI@POC Team is designed for physicians and healthcare system members seeking to leverage augmented intelligence to improve patient care, clinical workflow and system operations. →Learn more

5. DATA STORAGE, ACCESS AND SHARING

KEY POINTS TO HIGHLIGHT IN THIS SECTION

Data access and sharing is a major component of data storage. Metadata and data tagging are essential components of effective data sharing. Data tags need to be standardized, optimized, and periodically re-evaluated.

Discussion

Data storage is one of the basic capabilities in a company's technology portfolio—yet it is a complex discipline. Most IT organizations have mature methods for identifying and managing the storage needs of individual application systems; each system receives sufficient storage to support its own processing and storage requirements. Whether dealing with transactional processing applications, analytical systems, or even general-purpose data storage (files, email, pictures, etc.), most organizations use sophisticated methods to plan capacity and allocate storage to the various systems. Unfortunately, this approach only reflects a "data creation" perspective. It does not encompass data sharing and usage.

The gap in this approach is that there is rarely a plan for efficiently managing the storage required to share and move data between systems. The reason is simple; the most visible data sharing in the IT world is transactional in nature. Transactional details between applications are moved and shared to complete a specific business process. Bulk data sharing is not well-understood and is often perceived as a one-off or infrequent occurrence. Data access and sharing is a major component of data storage.

Metadata and data tagging are essential components of effective data sharing. Data tags need to be standardized, optimized, and periodically re-evaluated.

With the popularity of big data, the growth of business analytics and increased information sharing between companies, it is much more common to share large volumes (or bulk) data. Most of this data falls into two categories: internally created data (customer details, purchase details, etc.) and externally created content (cloud applications, third-party data, syndicated content, etc.). The lack of a centrally managed data sharing process typically forces all systems to manage this space individually, so everyone creates their own copy of the source.

As organizations have evolved and data assets have grown, it has become clear that storing all data in a single location is not feasible. It is not that we cannot build a system large enough to hold the content. The problem is that the size and distributed nature of our organizations – and the diversity of our data sources – makes loading data into a single platform impractical. Everyone does not need access to all the company's data; they need access to specific data to support their individual needs.

The key is to make sure there is a practical means of storing all the data that is created in a way that allows it to be easily accessed and shared. You do not have to store all the data in one place; you need to store the data once and provide a way for users to find and access it. Once data is created, it will be shared with numerous other systems; it is critical to address storage efficiently, in a way that simplifies access.

A key component of storage includes creating a common metadata layer, as it lets a company

consistently repeat the data preparation process. Common metadata also provides lineage information so you can answer such questions as:

- Where did the data come from?
- What quality attributes does the data contain?
- What data was used, and where else has it been used?
- How was the data transformed?
- What additional reports or information products are, or have been, developed using this data?
- What data elements do subject-matter experts typically focus on?

Applying metadata across the data analytics life cycle delivers savings on multiple levels. When a common metadata layer serves as the foundation for the model development process, it eases the intensely iterative nature of data preparation, the burden of the model creation process and the challenge of deployment. The outcome is more efficient collaboration, better productivity, more accurate models, faster cycle times, more flexibility, and auditable, transparent data.

The more, relevant data you can apply to a business problem, the better its potential solutions. While there is no shortage of data available to your enterprise today, it is often difficult to know what data is accessible and how it can be used. The ability of disparate data to connect and combine (even when it is co-located in the same data lake or cloud repository) is largely dependent on the metadata the data shares. Data tagging is only one aspect of that—but it is a very important one.

Within the context of enterprise data management, data tagging provides many benefits. For example, data tagging can:

- Help determine how much data preparation should be performed on new data sources.
- Enable efficient data discoverability—so when data is needed later for specific business purposes, it is a quick and easy process to locate the most applicable data.
- Improve big data quality, especially by making unstructured and semi-structured big data more usable.
- Help identify sensitive personal data so access can be properly managed and governed.
- Help flag and filter ethically dubious or otherwise questionable data before any of it is used in decision making or artificial intelligence solutions.

Guidelines for Data Tagging

Good practices for data tagging are presented below.

1. **Standardize the Tags**
   Data tagging is a subset of the essential metadata that makes up a business glossary. The business data term list in a business glossary provides an authoritative vocabulary that promotes a common understanding between stakeholders in an organization. Without establishing standard values, tagging often produces homonyms (i.e., the same tags used with different meanings) and synonyms (i.e., multiple tags for the same concept). This variability can lead to inappropriate data relationships and inefficient searches for data regarding a particular subject.

2. **Use All Applicable Tags**
   As with a lot of metadata management tasks, the user might get some results by doing the bare minimum by only applying one or two tags. But since most data can be used for multiple purposes, it is important to use all applicable tags. Doing so may reveal unexpected uses. It could also identify business groups most interested in tagging a particular source – which might make

employees in that functional group a logical owner of data stewardship.

3. **Do Not Over-Tag**
   This recommendation sounds like a contradiction of the previous recommendation. But tags can lose their significance through overuse. Frequency distribution analysis of tag values, both individually and in various combinations, can help prune extraneous tags for optimal efficiency. This analysis may also help further standardize the tags by revealing that a frequently used combination of tags should be available as an additional standard tag value, which is sometimes better suited than assigning multiple, individual tags.

4. **Re-evaluate Tags Over Time**
   It is important to remember that business terminology and business context rarely stay static. While many tags remain applicable for a long time, do not assume they will always stay that way. Plus, if tagging does not consistently deliver the benefits described above, then investigate why. Users may need to re-standardize and re-apply tags.

6. DATA STREAMING

KEY POINTS TO HIGHLIGHT IN THIS SECTION

The Internet of Medical Things will bring major patient benefits and rely on data streaming. Streaming sensors and algorithms can perform real-time diagnoses and patient monitoring, in place of diagnostic tests in the physician's office. Data security is a significant issue. Having data processing as close as possible to data source reduces data streaming and boosts performance. Data streaming from a sensor needs to be monitored and cleaned in real time.

Discussion

Streaming data is data emitted from a device and in motion, not persisted in any permanent storage mechanism, and can include two-way streaming. It is high-volume and high velocity. It flows into organizations from satellites, sensors, meters, phones, the internet—from all kinds of devices—and it comes in many different forms. Streaming data is transmitted continuously from virtually any source at rates of millions of events per second.

Data is arriving faster, and more unstructured when one considers image and sound. This is possible now as wearable devices are getting cheaper and increasingly used, based on lower power needs, and deployed in areas they were not deployed before due to the continual advancement of high-speed internet. The many streams of data can now be analyzed to create models that can predict specific outcomes. For example, one could predict asthma attacks from analyzing breathing patterns.

With the rapid growth of IoT with embedded AI (e.g., machine learning), organizations across industries are now dipping their feet into streaming analytics to: 1) Deal with large volumes of never-ending streams of events; 2) Manage low-latency applications to gather insights as fast as possible; and 3) Process data at the network edge when there may or may not be always-on connectivity to transport the data back to the cloud (or, enterprise core).

Factors to Consider

Working with IoT data involves unique considerations relative to data quality. Practitioners often discuss the following as challenge areas: a) Imbalanced and high-frequency data, b) Low signal-to-noise ratios, c) data not fully exploited—less than half of structured data is actively used in decision making, and d) what data should be stored versus what data should be sent back from the edge?

When it comes to working with IoMT additional considerations exist: a) determining objective of analysis and acceptability of digital endpoints b) security and privacy of device and data; and c) data strategy—how is the data collected, cleaned, analyzed and stored?

Determining objective of analysis and acceptability of digital endpoints

In working with streaming data analytics, just as in other data analysis, the first critical step is to define the objective. Once the objective is clear, the next step is to identify new clinically relevant endpoints which are easier, faster, and cheaper to track and more convenient for the patient to measure and act upon. Then, one can think through a way the collected data (e.g., motion and balance data) can be validated and correlated to medical outcomes data. The medical outcomes data will be then be used create an algorithm and produce results that can be used for informing medical decisions.

To illustrate this approach, say for example, one wants to measure arthritis using sensors. The traditional approach would be to put a patient through a series of 14 simple physical tests called the Berg Balance Scale, that takes 20 minutes of a physician and patient time, and is quite subjective. Could small sensors measure movements, delivering gyroscopic and accelerator data, combined with machine learning algorithms, deliver more useful quantitative data, making the Berg Balance Scale a quantitative, digital, automated clinical assessment that can be reproduced? In order to do this, Dr. Mark Wolff from SAS collected data from a variety of sensors and Intel processing, and created mathematical models to describe the preoperative condition of the patient as recorded in the assessment—a quantitative baseline. From there, models are used to score and track rehabilitation progress and, ultimately, to evaluate the effect of surgery in terms of therapeutic efficacy compared to the primary diagnosis. While being led through the Berg Balance Scale assessment, the system recorded a host of measurements—far more than any clinician can measure. This data was fed into a model that could ultimately be shared across providers treating others affected by arthritis. Subsequent routine assessments are fed into the same model, which could allow doctors to track the patient's progress with pinpoint accuracy in rich detail, seeing patterns that would have otherwise remained hidden. Using machine learning, with more data, the models themselves improve over time.

For more formal studies, there is no guidance with respect to the use of digital endpoints in a study, therefore the FDA's guidance (2009) on Patient Reported Outcomes (PRO) endpoints can be used as a foundation. The lack of clear guidelines and standardization across countries and regions has led to cautionary usage of digital health technologies to generate digital endpoints in clinical research.

Identification of suitable outcomes that are meaningful to patients and align with the overall study objectives, a device can be selected to capture those data under consideration. Sponsors will need to develop a strategy to ensure that the digital endpoints are acceptable to the regulators. In that same way, all data submitted to regulators need to meet minimum standards in terms of validity, reliability, sensitivity, and robustness. Agencies will require a similar evidentiary dossier to support the use of data from a specific device in any given study. In addition to the use of devices to generate digital endpoints, other required evidence includes intra- and inter- device variability, methodology of ensuring patient compliance with a device, and details of the patient support tools that will be put into place as patients go on their digital journey.

Security

Securing data while in motion through encryption from device to cloud is critical to maintain and assure data quality. Medical devices and their users also may share data across various operating systems as there are no standard operating systems for medical devices. There is need to secure the device from tampering with data that is on the device as well as data while it is in motion.

Generally speaking, a key best practice for machine learning is to achieve minimal data movement—a goal is to push data processing down to the data source to dramatically boost performance.

Data streaming strategy

Critical to managing the high-volume, high-velocity data sets is to establish a data strategy upfront, which includes the following considerations: data cleaning/data standardization, data continuity, data veracity, and data ingestion[10].

[10] Definition: "Data ingestion is a process by which data is moved from one or more sources to a destination where it can be stored and further analyzed." Source: https://www.alooma.com/blog/what-is-data-ingestion

When it comes to medical device-based sensors, data quality needs to be high and consistent, and there needs to be processes in place to ensure data standardization and on-going quality. For example, you do not have dirty data in the traditional sense of user entry of data when dealing with streaming data, but more from how the device is used. Traditionally, low-quality data comes from use errors during data entry, however poor quality streaming data is often a result of how the device is being used (e.g. a wearable on the wrist that estimates the number of steps the user takes in a day can under-report if the user is pushing a grocery cart because the arms are not swinging as they typically do when walking.) It is quite easy to determine if poor data is being produced, since it can be compared to clear target numbers that one would expect (e.g. blood pressure, temperature.)

The other challenge when one collects more data at a higher frequency is that there is the potential of generating unrelated alerts that result in alert fatigue (the user stops responding, since they assume the alerts are irrelevant). Therefore, there is a necessity to enhance automation and intelligence to check quality and consistency of the data through data veracity – How does the device algorithm generate accurate, precise, interpretable, and trusted outputs?

Involving subject matter experts through the algorithm training process is important for interpretation and context. Additionally, the subject matter expert can identify alert and action limits, as well as erroneous data For example, the algorithmic system for a blood pressure monitoring device will include blood pressure limits with an auto-rejection system to exclude invalid data from the analysis, such as negative blood pressure readings. The automation training process will need intelligence for the system to make that judgement. One way to accomplish this intelligence is for the algorithm to assess data across different sensors in order to auto-compute normal situations/data in which a certain endpoints would have been expected. For example, detection that the person stood up and got dizzy, so the blood pressure was affected makes the alerts 'smarter.' The connection across multiple data points can correlate alerts to relevant risk factors (e.g. heparin anticoagulation therapy must be monitored very carefully[11], and the algorithms can be set to be more sensitive, as Heparin has a narrow therapeutic window.)

Data continuously received from devices must be processed in a similar manner to ensure consistency in data quality. It should also be noted that streaming data (e.g. from wearables) could be cleaned in real-time, which eliminates lag, and reduces time to identify any issues. Algorithms can increase the level of sensitivity of the device to detect anomalies or radical increases/drops in measurements (such as temperature), and distribution changes outside of normal bounds.

For some applications with limited communication bandwidth (wearables being just one example), once the anomaly or malfunction is identified, it may make sense to NOT send erroneous data at all. In the event that further investigation is needed, (e.g., to determine if the patient is getting sick or if the device is malfunctioning), then a full set of data can be streamed to identify the issue and allow for external analysis of the full data set.

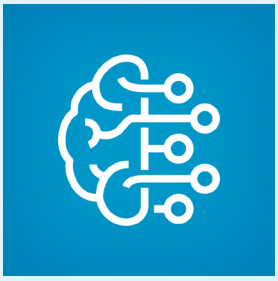
Benefit: Given that much of health data can be bracketed with rules, the anomaly scoring method allows for an exception-reporting strategy, which reduces storage costs and volume of transfer data.

There are techniques that are very helpful in identifying the most relevant data to use, including:

[11] https://techtransfer.universityofcalifornia.edu/NCD/27178.html

- RPCA (Robust Principal Components Analysis): used to determine the most valuable data to include for training models, and identification of what data tags are relevant and significant for predictive capabilities.
- Tools that help separate signals from the noise are very helpful.
- Using algorithms that allow for various operational scenarios (algorithms for sitting versus walking and linkage to dizziness for example) can generate more accurate predictive models.
- Identifying the appropriate algorithm to use for streaming and various data sets.

XAVIER HEALTH AI EFFORTS



AI Summit and Workshops

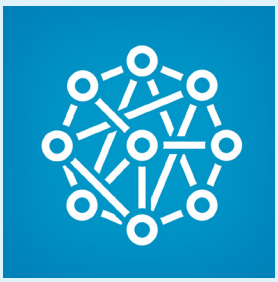
Advancing world health by bringing thought leaders and stakeholders tozgether to share opportunities, successful practices, and lessons learned that will advance the responsible and effective adoption of AI solutions.

→Learn more

AI Working Teams

Join a chartered working team to make a difference across the industry and patient lives around the world.

→Learn more



AI Experts Network

This network is intended to engage members in discussions on successful practices, challenges, and ideas around the development and adoption of AI.

→Learn more

XAVIER HEALTH

7. DATA RISK MANAGEMENT

KEY POINTS TO HIGHLIGHT IN THIS SECTION

Risk management for adaptive systems must be ongoing. Traceability and diagnostics built into the design will aid future failure investigations. Benefit-risk assessments should be performed before making upgrades.

Discussion

What is risk management?

There are typically two types of risk management—one type is focused on organizational goals, and the other is focused on safety.

Organizational risk management focuses on business interruptions due to natural disasters (floods, hurricanes, tornados, etc.), fires, and supply interruptions. However, there are risks such as the impact of a global pandemic that may not be considered in routine risk management planning. Traditional risk management typically includes risk to the business, buildings, and employees, and involves planning around known potential hazards. As became evident with the emergence of the COVID-19 pandemic, we are not always prepared to react to unknown risks.

Safety risk management considers the physical and clinical risk of the product to the customer once it is released for sale. Typically, the product risk process considers the functionality of the product—does it deliver as designed, what are potential malfunctions, are there usability issues, etc. The focus is on the potential for harm to people, property, or the environment.

Fortunately, both types of risk management follow a similar process—you identify goals (purpose), and you identify potential sources of risk, evaluate them, put risk controls in place where needed, assess residual risk, continuously monitor to identify new risks, etc. For AI applications, this same risk process can be used, but the potential sources of risk and the way failure occurs could be different than traditional software applications. This section will focus on the differences between traditional software risk management and AI risk management.

Factors to Consider

The quality of the development and test data obviously has an impact on product performance, with the potential of affecting the risk profile of the product. Data that is incomplete, incorrect, inconsistent, biased, etc., can lead to realized product or business risks. There could be factors that you are not aware of that could affect product performance (e.g. ambient temperature or humidity may be hidden factors not captured by manufacturing data). Risk analysis is used to identify and then minimize the potential number of surprises during execution of a project.[12] Although we cannot precisely predict future events, we can engage in risk management processes to identify risks associated with the application as defined, and secondary risks that may develop, based on decisions made during product development.

[12] Lavanya, N., & Malarvizhi, T. (2008). Risk analysis and management: a vital key to effective project management. Project Management Institute. Paper presented at PMI® Global Congress 2008—Asia Pacific, Sydney, New South Wales, Australia. Newtown Square, PA: Project Management Institute. https://www.pmi.org

Common tools include a risk register, risk checklist, and risk repository[13]. The nature of adaptive systems introduces extra complexity to the product design. Adaptive systems can learn over time, and the procedure for updating an adaptive system is critical. Software updates also raise cybersecurity issues which should also be considered. In adaptive systems, chaos theory applies since the system being developed seeks to simulate human behavior and decision-making. The challenge is to simulate systems that do not always have linear and predictable paths[14].

The black box nature of some AI applications may make it difficult to perform a root cause analysis when investigating issues that arise with the data and/or outputs of the algorithm. If the software is not performing as it should, is this due to a latent software defect? Or bias in the original training data? Or problems with a recent update to an adaptive system? Or some other cause? The design team should consider what traceability and diagnostic features should be added to help aid future forensic quality investigations. In developing the system, can you specify what it is trying to achieve or optimize? What assumptions can you make about the environment that the system is trying to simulate? In testing the system, what assumptions were made that became the boundaries or operating principles for the system in achieving the design goal?

Guidelines

This paper recommends the following practices:

a. A risk analysis should be performed regarding the data acquisition, cleaning, manipulation, and storage.

b. A risk analysis should be performed regarding the process used to update adaptive systems, including timing, training, potential errors during the update, consequences of update failure, etc.

c. Risk controls should be verified to ensure that they are effective. Additionally, an analysis should be performed to assess if the risk control itself introduces new risks (e.g. unintended consequences.).

d. Before making an update, the residual risk should be considered and a benefit-risk analysis should be performed that demonstrates improved benefits or reduced risks as compared to previous versions.

e. Diagnostic features should be considered to help in future forensic investigations

f. Product performance should be monitored post-market, looking for anomalies that might indicate missing data, bias, underfitting, overfitting, etc. as well as the integrity of data storage (e.g. has it been corrupted or compromised?) as well as changes external to the product that may impact performance (e.g. potential new data sources, changes in assumptions, etc.) There may be a need to involved subject matter experts (e.g. physicians) to help identify and track these external factors.

[13] Project Management Institute (PMI). (2020). Standards overview. https://www.pmi.org/pmbok-guide-standards

[14] Erik Erikson, "Introduction to Complex Adaptive Systems and Risk Analysis", https://www.ifpo.org/wp-content/uploads/2019/07/Risk-Analysis.pdf

8. UNIQUE REQUIREMENTS

KEY POINTS TO HIGHLIGHT IN THIS SECTION

Analytic data inputs are used for conducting various types of analysis including the validation of various business analytics. Data inputs and outputs, as either structured or unstructured data, should inherently contain specific attributes to maximize their utility. Analytic Data visualization and golden datasets are effective techniques that can be used to successfully test specific outcomes. Data for continuous improvement should be stable over time and contain the same attributes as the input and output data used in the initial analytic application or model.

Discussion

There are unique requirements that should be considered for data used in AI applications. These unique requirements apply to analytic inputs, testing/validation activities, data outputs and data used for continuous activities. The unique data requirements are essential to enable AI to function effectively and in a compliant manner.

Analytic Inputs

When looking at data for any AI use, the user may evaluate structured data or unstructured data. Structured data is data that is clearly defined through specific content and an easily searchable pattern. A company's employee roster stored as a relational database is an example of structured data. Unstructured data however is typically not easily searchable and includes examples such as videos, phone conversations and social media posts.

Analytic inputs are the various forms and types of data that are used for descriptive, diagnostic, predictive or prescriptive analysis.

Testing & Validation

Analytic inputs are used to test or validate outcomes and conclusions for various business analytics. These include descriptive, diagnostic, predictive or prescriptive analysis. There are, however, specific data requirements used for testing and validation activities that users should consider.

One key activity related to testing and validation is data visualization. Data visualization allows the user to see a graphical representation of the data to gain certain insights into patterns and trends. These patterns and trends are tested and validated with subsequent analysis. For example, plotting boxplots of specific manufacturing data, per machine setup, over time allows the user to understand whether the process average or variation is shifting over time. This is a critical analysis to determine if the data is stable over time and can, therefore, be used as an accurate predictive model. Another example may be a scatter plot comparing pH and viscosity. The user may want to conduct this visual testing prior to conducting the formal validation of the predictive model on the correlation between viscosity and pH. Figure 1 provides a graphical representation of the relationship between two variables that would help the user confirm the initial assumptions of the data set during the test phase.

Another tool used to validate data visualizations is "golden datasets". These datasets are especially helpful when looking at descriptive and diagnostic analytics, and can provide the user with a known and expected output to help determine if a change in the output is related to the data flows, model, business logic, etc.

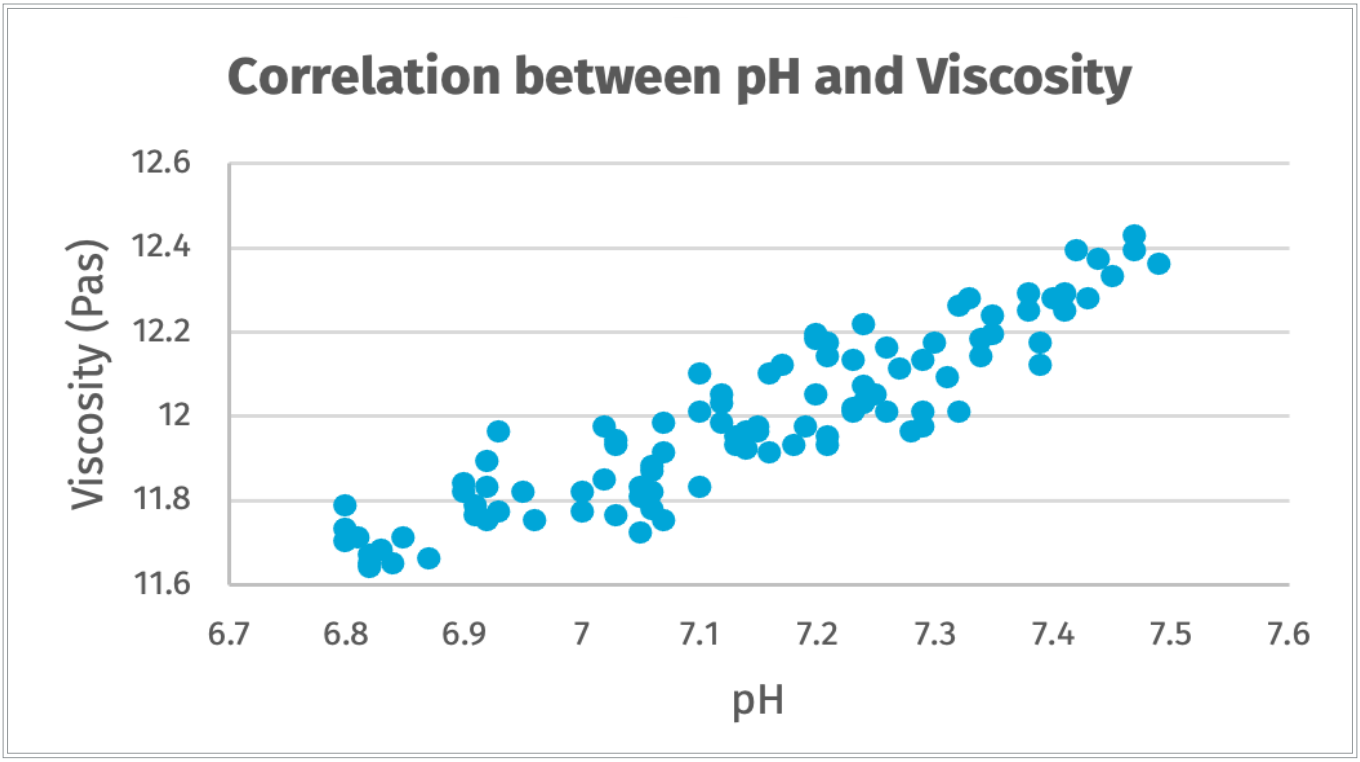


Figure 1

Data Outputs

Data outputs used in any business analytic application have unique requirements. A data output is the outcome of specific data inputs and may come from various sources. The data output may be data that is obtained from a particular data stream/data field, or it may be part of an analytic to be calculated based on the intent of the predictive or prescriptive analytic model.

For example, the user may collect manufacturing data that contains the pH and viscosity data and establish a specific correlation to characterize the relationship between the two data sets, or the user may use the pH and other process data to predict the viscosity via a calculated function or mathematical model.

Business applications often involve the development of specific algorithms. There are certain practices that could be applied to build, manage, and monitor the ongoing health of the algorithm. These include the performing ongoing verification and/or periodic health checks of the algorithm.

Data for Continuous Improvement

Once the analytic data inputs have been collected, the testing and validation data determined, and the data outputs assessed, the user should evaluate the data for continuous improvement. The continuous improvement data is essential to continue to add value to the business, agnostic of the analytic use (for example, descriptive or predictive).

Factors to Consider

Analytic Inputs (structured data)

- Data completeness
  - » Does the data contain data attributes?
  - » Is the data set complete or is it a sample of the data set?
  - » Does data or attributes need to be combined to provide better meaning?
  - » What categorical values does the data contain and how are they represented?
  - » If required, is the data time-based?

- Data Propriety
  - » Is the origin/lineage of the data known and can it be used?
  - » Can the data be trusted?
  - » Are there any legal issues in using this data?
  - » Is the data regulated or restricted in some way?
- Data accuracy
  - » Is the data historical and factual?
  - » Is it representative?
- Data errors
  - » Can data errors be detected?
  - » Are errors present?
  - » Can errors be corrected?
  - » What is the error frequency?
- Missing data, null/empty values
  - » How much data is missing, and does this render the data "incomplete" ?
  - » Are any data attributes missing?

Analytic Inputs (unstructured data)

- How are the documents or content organized?
- Are there certain areas or locations of the content that contain more data than others?
- Does the content or documents need to be divided into smaller components?
- Does the content contain data formats such as tables or other graphics that need to be deconstructed into actual data?

Testing & Validation

- Data visualization
  - » Does the data (inputs or outputs) contain any anomalies or unexpected irregularities relative to shape, center, spread, outliers?
  - » Is the correlation between the data output and data inputs visually apparent?
  - » Is this correlation stable over time?
- Golden datasets
  - » Does the known data validate the expected outcome of the model?
  - » Is the relationship between the data flows, model, business logic, etc. to the data input (through the Golden dataset) maintained?

Data Outputs (structured or unstructured data)

- If there is missing data, can it be corrected/ updated?
- Can the data source be trusted (in cases where the output is obtained and not calculated)?
- Is data representative of the process or source that generated the data?
- When data is plotted versus the data input(s), does it follow the expected pattern based on initial assumptions?
- Does the data need to be normalized, stratified, or transformed in some way to better relate to the data input(s) and/or model in general?
- Is the model appropriate for the data output (for example, categorical inputs trying to predict continuous data outputs)?
- Is the level of granularity/resolution between the data inputs and outputs consistent or does it need to be modified?
- Does the data contain the right level of detail to provide meaningful insights?
- Can the strength of the correlation between data inputs and outputs or model effectiveness be adequately calculated?
- How are outliers or data anomalies going to be identified, investigated, and utilized in the overall model?

Data for Continuous Improvement (structured data)

- Data completeness
  - » Data attributes/categorical values have not changed such that the analysis, model, or output is not altered

  - » Data continues to be time-based as required
- Data Propriety
  - » The origin/lineage of the data has not changed, if it has changed, the appropriateness and use has been confirmed
  - » The data can continue to be trusted
  - » There are no outstanding legal, regulatory or restrictions on the use of the data (if there are any, there is a mitigation plan and strategy top address)
- Data accuracy
  - » Factual origin of data has not been altered
  - » Ongoing data continues to be complete and representative
- Data errors
  - » The error rate of the data is consistent or better than the original data set validation
  - » Potential errors can be corrected and mitigated
- Missing data, null/empty values
  - » Appropriate data attributes continue to be present
  - » The rate of missing data is consistent with or better than the original data set validation

Data for Continuous Improvement (unstructured data)

- Source documents and/or content continues to be organized the same way
- The locations/areas of the content that provide critical data is consistent and relatively stable over time
- The deconstruction of data for a given content or document does not need to be altered on an ongoing basis

Guidelines

Analytic Inputs

A key take-away for analytic inputs is that data in and of itself has no value, so the user must ensure that the data is relevant and contains the proper fields and other metadata attributes and descriptors that enable business decisions. The user should consider running basic statistical analysis such as mean, minimum/maximum values and standard deviation on the inputs to ensure that the data makes sense. For example, if the data represented process parameters that span a certain range, say 3-200 units, it would be expected to contain decimal values (for example, 0.231 units).

Testing & Validation

The user should recognize that there will be specific challenges given the type of data (structured vs. unstructured). The testing and validation practices here will help mitigate the potential error in subsequent analysis and model maintenance. Most of the data input requirements previously described will provide benefit and value in the testing and validation of the data.

Data Outputs

Analytic data outputs are like analytic inputs in that it should contain the proper metadata attributes. Furthermore, analytic outputs result from, and should correlate to analytic data inputs. This correlation is established through various analytic tools and is maintained for the duration that the relationship between inputs and outputs is maintained.

Data for Continuous Improvement

Continuous improvement data should be stable over time and contain the same information in the same way as the input and output data used in the initial analytic application or model.

9. CONCLUSION

This paper was developed by the Good Machine Learning Practices Team in Xavier Health as a follow up to the whitepapers published in 2018 and 2019 regarding good machine learning practices[15] and explainability[16] "Perspectives for Good Practices in Continuously Learning AI Systems in Healthcare" published in August 2018 and the "Building Explainability and Trust for AI in Healthcare" in 2019. As the title indicates, our focus here was to establish a discussion around factors that impact Data Quality for AI applications.

Machine learning systems do not just extract insights from the data they are fed, as traditional analytics do. They change the underlying algorithm based on what they see in the data. The more relevant data they are fed, the more tightly they define the algorithm and the more confidently they make classifications or predictions.

The "garbage in, garbage out" truism that applies to all analytic pursuits is truer than ever. If the data that feeds machine learning algorithms is not well managed, the results could be like the result of the whisper game – wrong statements where errors have multiplied upon themselves. The following dangers are obvious: inconsistency, inaccurate insights, loss of trust and AI results becoming stale.

This paper outlines the many aspects of Data Quality that are important for successful and sustainable implementation of AI in healthcare. This paper addresses the critical elements for Good Lifecycle Data Quality management (GLDQ), including characterizing the data to be acquired, potential bias in the data, data selection, cleaning and manipulation, and storage. Additional topics such as data streaming, risk management, etc., were also discussed.

It is clear that data quality has a direct impact on product quality and performance, and we hope that this paper has helped in your understanding of the topic.

[15] Baird, P. Nayak, R. et al (2018) Perspectives for Good Practices in Continuously Learning AI Systems in Healthcare, Xavier Health & Xavier University

[16] Baird, P. Nayak, R. et al (2019) Building Explainability and Trust for AI in Healthcare, Xavier Health & Xavier University

XAVIER HEALTH

Inspiring collaboration. Leading innovation. Making a difference.

www.xavierhealth.org