The Tsunami of Big Data for Pharma: Sink or Swim?

This article addresses the real-world challenges in assembling Big Data that need to be considered when developing and applying analytics to enhance drug and diagnostic development and patient management. AI/ML and deep learning tools focus on volume and velocity but the real value will come from understanding and dealing with aspects of validity that are currently being ignored.

Michael N. Liebman
PhD, Managing Director
IPQ Analytics, LLC

The concepts of Big Data and Big Data Analytics have been around for some time but it has only been since the late 1990's, with the confluence of genomic and transcriptomic data, along with increased use of EHR's and access to claims data, that Big Data has arrived at the shores

> Embracing big data offers a competitive advantage, allowing companies to adapt to evolving market trends and patient needs.

of pharma and healthcare. This rapid and accelerating access has been frequently represented by Hokusai's tidal wave (Figure 1) but this metaphor may be hiding the real challenges that we need to face to deliver more effective diagnoses, drugs and patient outcomes. A potential evolution of this model, to embrace the critical complexities of Big Data, can be accomplished in the transition outlined in this figure…as a "leaning tower" of books and publications, etc and outline in this article some of the "critical challenges that exist in the details".

The size, the power and the non-predictable nature of the tidal wave metaphor served as an early warning to the healthcare and life sciences communities that major changes were imminent. This has led to technological development and implementation in hardware (e.g. massively parallel computing, GPU's, quantum computing), in software (e.g. AI/ML, deep learning, generative AI) and in "cloud-ware" to handle two of its "V's", the volume and velocity aspects of big data, but has not necessarily focused on its third, validity. With the increasing recognition and utilization of complex analytics to interpret Big Data, an emphasis on validity is critical but needs to be expanded beyond accuracy as currently defined for Big Data.

The tower of books metaphor can readily point to several realities that map to real world issues in Big Data analysis:

1. While many new books are published each year, there can be significant differences in quality between those which are self-published vs those that have gone through a traditional review and publication process;
2. Each book reflects the story that the author wants to tell, some more truthful and some more fiction, and likely none without some bias and incomplete rendition of events, both sides of the story;
3. Books are published in many languages and their translations may not accurately convey the intent of the author as expressed in their native language e.g. idioms;
4. Scientific, technical and medical books are published focused within disparate disciplines where use of specific words

may have different meanings among those disciplines;

5. Books are written in different formats, e.g. novels, dictionaries, instruction manuals, poems, etc;
6. Books have different numbers of pages, words and figures;
7. Books are commonly written to describe specific periods of time related to the story;
8. Books typically reflect the state of knowledge and use of terminology pertinent to a specific time period and both may be subject to change over time
9. It is also worth remembering that "you cannot judge a book by its cover" and that extends to databases as well (as detailed below).

Evolving the tidal wave to the leaning tower of books provides a perspective on validity that highlights critical challenges and constraints that may exist in Big Data and significantly impact its subsequent analysis, interpretation and usefulness in drug and diagnostic development and patient management. As evidenced during the last two years of the COVID pandemic there has been an explosive growth of books, papers and data concerning COVID, well beyond the current capacities of the scientific and clinical publication systems, resulting in limited validation even at the journal review level. It is estimated that within the next 10-12 years, the number of scientific journals and the number of scientific publications will double the current totals. In evolving the metaphor, the tower of books, i.e. databases, is more representative of the actual challenges to aggregation and integration of data from disparate data bases that comprise Big Data and which are more significant than the increased volume, alone.

Within each component database, several considerations include:

1. New databases are being created (or expanded) each year:

a. Each database reflects authors/creators/curators who provide their own perspective to determine what and how data is collected and stored, i.e., emphasizing a particular clinical or experimental specialty, e.g. radiology, pathology, gene expression/transcriptomics

b. Data collected commonly results from ease of access rather than attempting to completely populate an objective model that comprehensively addresses the problem e.g. patient journey (pre-disease to outcome), process of diagnosis.

Evaluation of gaps within social determinants of health and their potential impact on clinical practice, health and research is intended to help reduce inequities in health among disadvantaged populations. Consideration of an individual's "zip code" or "census tract" is used as a surrogate to evaluate socio-economic factors, environmental exposures, educational access, etc. However

this does not adequately model the reality that an individual's daily activities, e.g. work, may present additional "exposures" on a daily basis because of the complexity involved in monitoring and integrating such activity. Additionally, cultural differences among population groups may yield significantly different prioritization of factors that comprise SDOH and result in very different responses to efforts to close such gaps.

2. Biases within a database can result from:

a. Populating a data model that is incomplete, inaccurate or biased, resulting in missing critical data;

Randomized clinical trials, which serve as the highest level of evidence in evidence-based medicine, establish inclusion/exclusion criteria that commonly establishes a trial population that does not reflect the complexities of real world patients who have comorbidities, poly-pharmacy, etc. or exclude significant population groups, e.g. no pregnant women were included in COVID vaccine trials (or many others).

b. Using an accurate model but having incomplete or missing data;

Missing data is a common occurrence that is sometimes handled using imputation, but this assumes a model is valid to generating the missing data. Separately, we use surrogate measures in place of complex physiological parameters, e.g. hypertension based on blood pressure monitoring. Episodic measures do not adequately consider concurrent factors, e.g. time to rest, extant stress, meals, etc, nor the normal diurnal variation that may be more significant in its variation than the single measurement over time.

c. Inadequate specification or definition of data fields;

Different algorithms maybe used to compute specific variables like Glomerular. Filtration Rate (GFR) where >5 separate algorithms are in common use, two of which incorporate factors that consider race of patient. Additionally a patient's diagnosis may be the result of the application different clinical guidelines and physician experience, none of which is noted. ICD-10 codes do not address this adequately as they are intended primarily to justify patient management and for reimbursement purposes.

d. Using different tests or test reagents to measure a clinical laboratory value;

Her2/neu is an epidermal growth factor that is over-expressed in some breast and other cancers and serves as a specific target for therapeutic intervention. The FDA has approved tests using immuno-histochemistry. (IHC), i.e. anti-bodies, to detect expression but studies have shown differential response to anti-bodies raised to different features of the protein. Additional test using in situ hybridization (ISH) detect gene copy number differences that can lead to over-expression. This also extends to instrumentation and different on site procedures for maintenance and calibration, leading to the need to ideally

use centralized facilities for multi-center trials, etc to minimize the variability.

e. Using different thresholds (standards) to assign results into either "+ or –" classifications;

Triple negative breast cancer (TNBC) is characterized by "negative" scoring in 3 tests: for progesterone and estrogen receptors and her2/neu (as noted above). Thresholds for a "+ or –" evaluation may vary among cancer centers and thus a TNBC patient may not receive the same diagnosis at different centers. Most recently, "+ or –" has been expanded to consider "low her2/neu" expressing patients further suggesting that specific values be used for clinical decision making and research rather than "+ or –".

f. Consider of temporal biases;

Potential temporal biases may develop from two different sources: One may be the period of coverage of a given study and resulting database, i.e. studying the effects of a drug on pregnant women and their offspring typically considers preterm births, birth defects and initial postpartum period (1 year), but developmental processes may not reveal impact until the child is much older, e.g. might SSRI's used during pregnancy impact neurogenesis and synaptogenesis and not show effects until adolescence with learning disabilities or behavioral issues. The second might reflect the external factors that were present when the study was done, i.e. changes may occur in standard of care, diagnostic criteria, testing procedures or interpretation, etc that could impact the data within a specific data base and, perhaps more significantly, when multiple data bases are being integrated or federated for analysis (see below)

3. Additional considerations concerning data validity:

a. Fit for Purpose: as noted above, all databases are initially developed and commonly maintained to reflect the needs and intent of their authors/creators/curators with some also evolving to serve the expanded needs of their user communities. In Big Data, the focus on aggregating or federating large data repositories has led to accessing most readily available data to "feed the analytic

engine" most typically involving AI/ML or deep learning, or to provide statistically significant power to the analysis. It is critical, however, to recognize the purpose for which the data was generated and collected. In real world data (RWD), the preponderance of data exists within "claims" databases rather than clinical records which may be more highly secured for regulatory and privacy concerns. Claims data may vary significantly in terms of its accuracy in representing the actual pathophysiology of the patient because of its potential use for justification of patient management and reimbursement. Where private insurance is predominant, e.g. the US, claims data is most representative of the "business of healthcare". Where healthcare is provided as a national service, claims data may more closely describe the patient's underlying conditions when justification for reimbursement may be less critical.

b. There is increasing use of natural language processing to extract additional data from clinical notes. Clinician's notes are not standardized, naturally reflecting the individual clinician's patterns/expressions and entered as needed. An additional "feature" has further confounded the use of clinical notes as most systems have incorporated the ability to "cut and paste" clinical notes to expedite physician entry, assuming that editing to reflect current evaluation will be made and that leads to potential duplication or carrying over of notes rather than updating and clarification.

c. Natural language processing of published articles and reports can also present challenges as some studies only include documents that are readily accessible, e.g. using abstracts in place of full text because of free access (PubMed) vs paywalls, and also accessing publications that are "self-published", i.e. early access to articles that may be in journal review but have not completed that process, and some which may never be accepted. Extraction of data and concepts from publications also should differentiate among the sections of the articles, i.e. data and methods sections and results section should be considered fundamentally more reliable than the discussion and conclusion sections where the author's interpretations are provided and may exhibit less objectivity.

Big Data typically refers to the aggregation/integration or federation of individual databases whose challenges are outlined above. In addition, there are enhanced needs for security and privacy concerns and regulations to be appropriately managed, national and international regulations for data exchange, compliance with disparate informed patient consents as to any limitations on personal data use and intellectual property considerations of analytic results. The distinction (and value) between de-identified data and anonymous data is significant especially when potential commercial products might result. Major efforts are underway at the national level, at the EU level, in the US, and within and across professional societies to

establish standards to support data exchange, e.g. FHIR, but these are mainly operational and may not address many of the underlying issues outlined above. In addition, while development, implementation and compliance to standards is laudable, it is also a long term process and will not necessarily address the current base of legacy data. It is critical to use this legacy data, with appropriate recognition and accommodation of its biases, etc to impact both research and clinical decision making now and using it to form the basis for the more standardized data future to which we aspire.

None of the issues raised here invalidate the potential use of the data for analyses, but they highlight challenges and constraints in the interpretation of the results.

No Big Data set will ever be perfect and complete. This reality provides both challenges to using Big Data but also presents opportunities to attain greater confidence in the results through incorporation of transparency in what the component databases and data actually represent. Segmenting and analyzing data provides a cascading approach for progressive addition and validation of data that may contain some of the biases noted here.

We typically use metaphors to convey complex concepts and make them more readily identifiable and relatable to a potentially varied audience. While this is well-suited to introduce new ideas, to realistically put these concepts into practice, it is necessary to acknowledge the real-world complexity of the problem/challenge/situation/process. This does not mean that all issues must be resolved to make progress, e.g. integrate Big Data for meta-analysis, but it does require a greater degree of critical thinking and planning. Effecting greater transparency as to these potential challenges to real world use of Big Data can greatly impact the validity, not only of the data itself, but also the accuracy of analytic analyses and interpretations that the data.

The opportunity for Big Data seems to be to "sink or swim"...to sink if these challenges create too many waves for comfort or to swim by adjusting to the real world nature of the sea. In dealing with Big Data it is sometimes worth remembering the quote of Mies van der Rohe that "Less is More". ■

AUTHOR BIO

Michael N. Liebman is currently working as a Managing Director, IPQ Analytics, LLC, has experienced both an academic and pharma/diagnostic career at Mt. Sinai, UPenn, Vysis, Wyeth, Roche where he has led programs and teams in Bioinformatics, Pharmacogenomics, Computational Biology, Cancer Biology. He has had senior advisory roles in PhARMA, HIMSS, IUPAC and leads IPQ in its international advanced analytics as a service (AAS) business in EU, China, Africa and Australia.