Leveraging Data Pathways for Next Generation Safety Monitoring of Medicines and Vaccines

Jeffery L. Painter
GSK
Durham, NC, USA
jeffery.l.painter@gsk.com

Laurie Girard
GSK
Durham, NC, USA
laurie.x.girard@gsk.com

Michael Glaser
GSK
Upper Providence, PA, USA
michael.x.glaser@gsk.com

Andrew Bate
GSK
London, UK
andrew.x.bate@gsk.com

***Abstract*—Safety evaluation of medicines and vaccines is critical to ensure patient safety and maintain confidence in treatment and disease prevention strategies. Leveraging data pathways for next generation pharmacovigilance (PV) requires the creation of new platforms that seamlessly integrate both structured and unstructured data. Here, we describe the design of a novel data environment that provides enhanced data mining, information retrieval, and data governance to improve PV processes and activities. The goal of which is to further inform the knowledge of potential safety issues during the life cycle of medicines, from routine healthcare delivery to informing future drug and vaccine development.**

***Index Terms*—pharmacovigilance, drug safety, vaccine safety, heterogeneous data, machine learning**

I. Introduction

Pharmacovigilance (PV) is the systematic and continuous evaluation of the safety of medicines and vaccines administered to humans during routine healthcare delivery. And, PV is paramount to ensure patient safety and to maximize our understanding from emerging issues during drug development and routine use in healthcare.

Diverse and increasingly vast amounts of healthcare-related data are transforming our approach to PV. Sophisticated data mining technologies, artificial intelligence (AI) and machine learning (ML) are increasingly capable of rendering large quantities of heterogeneous data into information that can be used to guide clinical decision making and drug development, and to help identify potential safety signals. PV has primarily focused on traditional data, such as spontaneously reported safety events of suspected adverse events (AEs). Including heterogeneous data requires a multi-modal approach to improve causality understanding and reveal potential safety issues relevant at the population level and for specific subgroups.

II. System Design

Fig. 1 illustrates a design that supports multi-modal data and includes data that are typically not well structured, but still contribute to the overall data ecosphere. Combining complementary data sources requires the development of novel processing methods and optimization for their use in PV.

Our design is a multi-modal safety data system and creates an infrastructure that will transform PV systems into a fully connected, modern platform. The figure illustrates that the foundations of patient data pathways are composed of multiple data sources (both traditional and non-traditional). Ontologies and knowledge graphs enable linking disparate data sets together allowing them to be utilized by more advanced modeling tools.

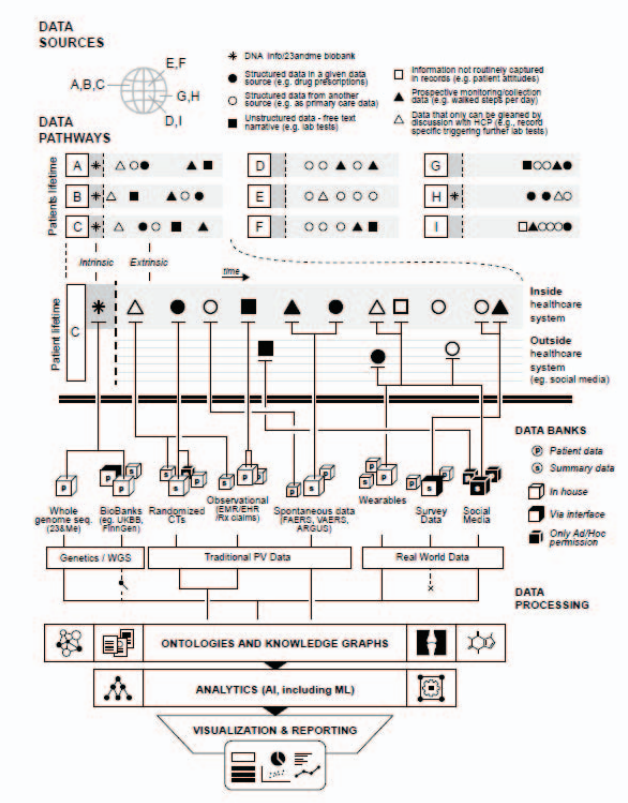

Fig. 1. Data pathways for next generation pharmacovigilance

This design also considers data access levels (summary versus patient level), privacy regulations, and the ability to link data together in meaningful ways. By enabling diverse data pathways, we have designed a modular system that is capable of efficiently dealing with issues as required (e.g. de-linking of data). Further, it supports data visualization to enable the rapid assessment and the ability to flag data issues (e.g. likely inferential safety concerns).

979-8-3503-2028-2/22/$31.00 ©2022 IEEE
DOI 10.1109/CSCI58124.2022.00321

III. A Modern Data Infrastructure for Pharmacovigilance

One of the strongest motivators for designing a new PV data infrastructure is the simple fact that traditional PV has relied primarily on association rules analysis for signal detection [1]. Most signal detection has therefore sought to improve upon these basic proportionality tests [2] [3] that do not fully take into account the diversity of disease, demographics and patient backgrounds. The low quality of data in safety reports may limit their usefulness, therefore making clinical review of quantitative outputs critical.

Furthermore, the language in which we express AEs in safety reports has its own constraints. Safety databases rely extensively on regulatory reporting requirements which mandate that medical events be expressed using the *Medical Dictionary for Regulatory Affairs* (MedDRA) [4]. The use of MedDRA has proven useful historically, particularly as a structured way for recording data as compared to free text. However, it is not without limitations. For example, MedDRA terms referencing genetic AEs are lacking, potential outcomes from immunotherapies have only been slowly adopted [5], and as an ontology, MedDRA is often misunderstood by the complexity of its underlying hierarchical structure [6]. Similarly, there are challenges with how drug or vaccines exposures are recorded and grouped together most effectively for analysis [7].

As we enable extended data pathways, attention has been given to the wider incorporation of systems biology, environmental conditions and disease pathways. These data linkages enable better approximations for identifying whether the data supports a causal link between medicines/vaccines and an AE, and represents a true paradigm shift from the historical, routine analysis of safety signals in medicines.

IV. Foundations for Enabling Data Pathways

A. Revealing the Layers of Data Accessibility

The first step in building our data environment is to focus on the various types of data that contribute to data pathways. To achieve this, our design must be agile in its ability to integrate each layer. The data sources can be viewed through the lens of data access and availability which include (1) centralized, in-house data, (2) remote access databases, (3) "ad hoc" use databases and (4) medical literature. Applying appropriate inference methods requires excellent understanding of each data source. Patient level data is anonymized to ensure data privacy is met, and each data source may also contain pertinent meta-data (e.g. data refresh dates, governance restrictions).

Historically, PV primarily focused on centralized, in-house data (e.g. safety reports, preclinical and clinical trials, pharmacology and *in vitro* models), and centralized data still covers a majority of PV needs. In recent years, PV has utilized remote access databases which mainly include real world data (RWD) with additional benefits. Occasionally, "ad hoc" use databases may be required which originate from various sources. These may or may not be used in day-to-day PV-related activities, but may be called upon as needed for more specific and detailed analysis, particularly of a suspected AE.

B. Centralized and In-house Data

1) Safety Adverse Event Reports: Postmarketing PV traditionally relies on the analysis of safety reports. Safety reporting is highly contextual and subject to bias, e.g. being more frequent when (1) a drug first enters the market, (2) the AE is perceived as serious, (3) the reporting environment is favorable, (4) attention is drawn to specific AEs by governments, media reports, or by litigation [8].

The absence of denominator data means that databases of safety reports cannot be used to estimate population-based incidence rates. Although linkage to drug utilization sources is sometimes done and can enable reporting rate estimation, this adds complexity to the analysis and is not an exact science (e.g. prescriptions are often left unfilled [9]). Duplicate reports and missing data are inherent problems that impact case counts. Missing data can render some voluntary reports uninterpretable in terms of diagnosis or causality assessment, and follow-up attempts for individual cases are often time-consuming and fruitless. Despite these limitations, postmarketing PV has been instrumental in identifying important or unexpected adverse reactions to vaccines and medicines that have led to label changes or, rarely, withdrawal of the product itself from the market [10].

2) External Safety Data Sources: Pharmaceutical companies collect, analyze, and share safety reports on their own products, but this approach to PV alone would certainly lead to a myopic view of the world. To support data pathways of relevance to PV activities, it is also necessary to incorporate all safety data that could potentially help link and connect suspected AEs within a broader context. To support this capability, our data pathways also make use of data sources that are well known in the PV community and readily accessible such as WHO[1] which is the global database of reported AEs for medicinal products.

C. Remote Access Databases

1) Real-world data: "Real-world evidence (RWE) is the clinical evidence regarding the usage and potential benefits or risks of a medical product derived from analysis of Real-World Data (RWD). RWE can be generated by different study designs or analyses, including but not limited to, randomized trials, including large simple trials, pragmatic trials, and observational studies (prospective and/or retrospective)" [11].

The term RWD is often used to denote specifically large, electronic healthcare records (EHR) or claims databases originating from multiple countries. These are rich sources of data containing millions of patient-level records suitable for longitudinal studies [12]. This type of data has long been used for epidemiological studies (which traditionally took years to execute – particularly when multiple databases were employed).

Now, RWD is increasingly used to investigate the natural history of disease, treatment patterns and outcomes, and specific AEs in relation to a drug or vaccine exposure. However,

[1] https://who-umc.org/vigibase/

use of these databases for signal detection in PV is still evolving and its value remains unclear (e.g. hypothesis-free signal detection). They also provide data not readily accessible through safety reports alone (i.e. determining incident rates of exposure, comparator drugs, demographics).

Limitations of RWD include data lag (typically 3–6 months) and biases based on database demographics. Additionally, RWD requires multiple analytical methods for different drug-event groupings. For example, capturing acute outcomes that occur shortly after exposure has different requirements than capturing deaths or diseases such as cancer that can take many years to develop [13].

2) Common Data Models: The use of common data models (CDMs) has made significant contributions to PV over the last 15 years. Development of data agnostic CDMs enabled one to perform comparable analyses across data sources by transforming the underlying data into a standardized format. RWD are typically not built for PV and CDMs are instrumental in transforming these data sources into a common format, enabling rapid data analysis and signal identification [14].

Our data environment can use RWD originating from countries such as the USA, UK, and Japan. These data sources are in routine use (Fig. 2), and our environment supports rapid inclusion of additional sources (i.e. those that are grayed out) which may be activated, near instantaneously, as needed in support of PV activities.

Currently lacking in the PV landscape is a harmonized, data format (native or CDM) agnostic, multi-stakeholder strategy that takes advantage of data systems as they emerge and evolve. For example, safety surveillance in resource-limited countries is frequently undeveloped or absent, representing a significant gap in terms of population representation in safety databases. To date, targeted strategies have been implemented to mitigate and encourage safety surveillance processes. However, these strategies are often specific to certain localities or outcomes [15].

D. Ad hoc Databases

1) Social Media Pipeline: Social media (SM) offers a non-traditional, worldwide data source that may be leveraged for PV activities. These data are readily accessible through aggregators or directly from the source (e.g. Twitter, reddit). In addition, SM data often contains geographic specificity, and provides direct access to the voice of the patient. There are also disadvantages including, not all data is accessible (e.g. Facebook restricts access at the aggregate level). The same AE may be duplicated on different forums. Further, SM users do not use standard drug names nor medical terminology of diseases and symptoms. It also contains high levels of noise (e.g. spam). Currently, SM data is not systematically used in PV and has been shown to perform poorly [16] [17]. However, it has been shown that analysis of SM data can have meaningful impacts on safety monitoring for some specific safety issues, such as enabling insights for patient engagement [18].

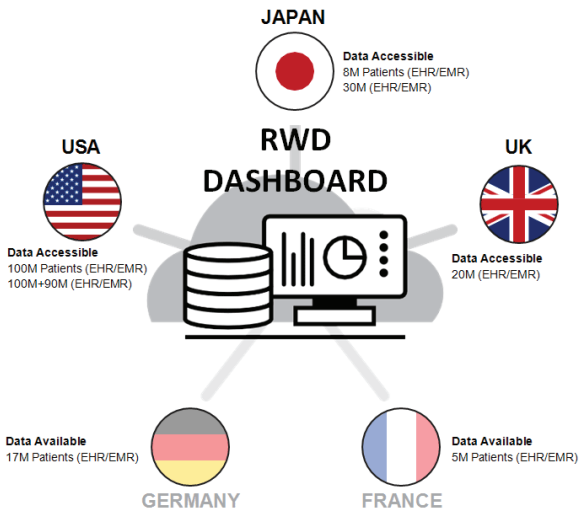

Fig. 2. Accessible and available real world data sources

We created an SM data processing pipeline as a module of our data pathway integration (Fig. 3). This pipeline is agnostic of channel source, automates the processing of unstructured posts using natural language processing (NLP), and defines a formal data structure that can be used for routine monitoring of suspected AEs.

SM data can be transformed into a format suitable for PV activities by following the data processing steps. (1) Standardization of drug names and mapping AEs to MedDRA; (2) data cleaning to remove duplicates, noise and spam using Bayesian probabilistic models (3) de-identification by removing personally identifying information. Once processed, the data is made available as part of our data pathways environment.

2) Systems biology and biobanks: Biomedical databases (biobanks) contain detailed genomic and/or health-related information, including medical imaging results, health outcomes and biological samples. Large population-based biobanks exist in several countries (e.g. Estonia, Finland, UK), but they tend to be concentrated in affluent countries [19].

Genetic polymorphisms among individuals exists within populations. Responses to medicines or vaccines, or even the risk of developing an adverse reaction to that product, are potentially variable among individuals receiving the same product due to genetic differences. Biobanks offer an opportunity to better understand why individuals may vary in their responses to a medicine or vaccine, and they offer the ability to define specific safety monitoring protocols for patients with higher risk. In-depth safety profiles may be generated by harnessing multifaceted datasets to help predict biological processes which may impact safety.

The SARS-CoV-2 pandemic has underscored the intimate connection between genetics, environmental data and disease, and their influences on holistic systems biology. For instance, the emergence of the COVID-19 virus uncovered the value

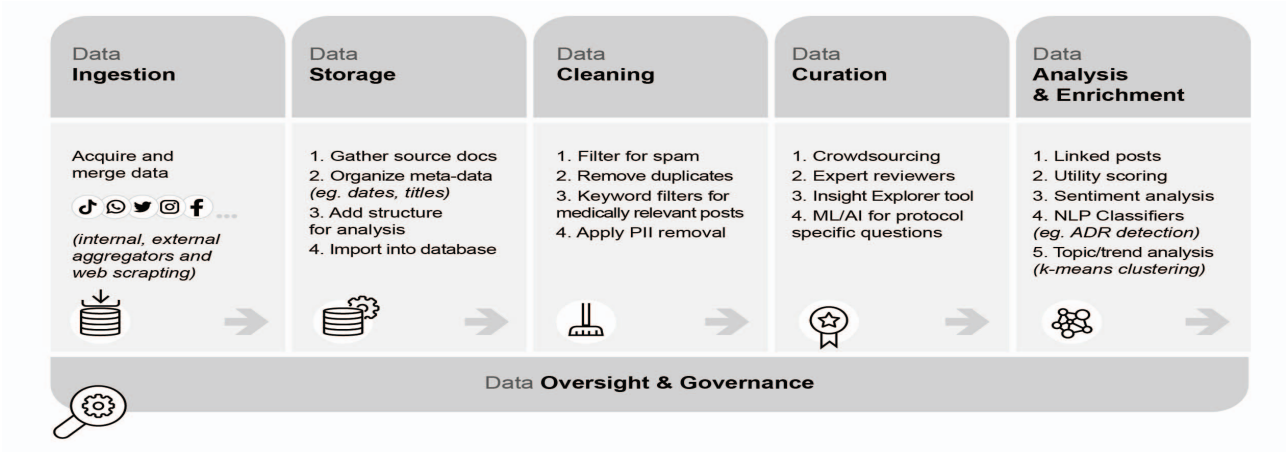

Fig. 3. Social media data processing pipeline

of "dark data", a term used to describe published data that is not connected to digital knowledge resources and is therefore unavailable for high-throughput analysis [20]. Liberation of dark data into digitally connected formats could expand research capacity and promote the development of non-linear outputs and re-use of data in relevant settings. Along the same lines, the use of biobank data and genome-wide association studies can provide potentially critical safety information that could contribute to a better understanding of the effects of drugs and vaccines on specific sub-populations; as concluded by Nogawa, et al. [21] in evaluating AEs associated with the COVID-19 mRNA vaccine.

3) Environment, weather and climate: Pollution levels, water quality, environmental exposures, natural disasters, pandemics, weather events and climate change all impact human health [22]. For example, during the COVID-19 lock-down, data collected from claims and EHRs (and other data sources) were potentially confounded, had missing data, or experienced additional bias due to an over-extended global healthcare system. Real-time monitoring of pollution and weather data can provide detailed information on these risks down to the level of a zip code. Such data can be used to complement safety signal detection and causality modeling.

Signal detection in PV does not tend to make any adjustments for where the patient is located and how various environmental conditions may affect their personal health journey. An individual's mortality is largely affected by where they grew up, the level of education achieved, and access to healthy food and proper healthcare [23]. Phelos et al. investigated nearly half a million trauma patients' lives and found that when vulnerability indices (i.e. Distressed Community Index[2] and National Risk Index[3]) were taken into account, these factors alone could account for determining outlier status due to geographic variation [24]. This can also help to account for variations in physical biomarkers, and when combined with emerging digital health technologies, could be leveraged to enhance the identification of AEs and potential causal associations [25].

[2]https://eig.org/distressed-communities/
[3]https://hazards.fema.gov/nri/

E. Medical Literature

In PV, the published medical literature serves multiple purposes, including (1) a direct source of safety data (e.g. safety reports, meta-analyses), (2) a reference when seeking to understand the mechanism underlying potential safety signals, and (3) provide background for benefit-risk review. Ad hoc searches on specific topics may also be requested by internal or external parties during signal investigation.

Generally, literature searches are carried out using platforms such as PubMED or EMBASE and titles are manually screened to identify potential articles of interest. As the number of relevant articles increases, particularly for legacy products, literature searches can become arduous and time-consuming. Manual reviews are also prone to error and individual reviewer bias. Narrowing search terms may decrease search results but comes with the potential loss of information. Other texts, such as media reports and gray literature, should also be searched, but they may be overlooked or missed altogether. ML and NLP techniques provide the ability to deal with large volumes of text and can improve the speed and accuracy of literature searches [26].

V. Innovative Approaches to Data Processing

Leveraging data pathways requires thinking more critically about systems management. Data linkage and retrieval must follow regulatory guidance and respect patients' rights, most notably, under the General Data Protection Regulation (GDPR). In addition, our data infrastructure platform aims to commit to FAIR (Findability, Accessibility, Interoperability, and Reusability) practices [27].

A. Data Processing and Accountability

On April 27, 2016, the European Parliament codified into law "the right to be forgotten", or what is now referred to as

the GDPR [28]. These types of edicts have now become law in other jurisdictions, although with potentially more limited applicability. Still, these types of data removal requests have real impacts in the global data ecosphere [29] [30] and must be taken into account in the context of a drug monitoring system. Our PV platform must include the ability to trace and audit changes in data. It must track and manage how the loss of data impacts both prior and ongoing PV studies, and it should alert users to when and where they can move data in accordance with regulatory requirements.

B. Data Enrichment

Data enrichment is the process of utilizing ontologies and knowledge graphs to add more value to data than exists in isolation. We have seen great success in the use of these methods in the annotation of human genetic data and the drug discovery process [31].

Eventually, we would also like to link PV data sources to biological pathways via the Kyoto Encyclopedia of Genes and Genomes (KEGG) [32]. Andersen, et al. suggested that gene expression can affect a patient's potential AE outcomes in their study on lymphatic filariasis, a neglected tropical disease [33]. The authors found a significant transcriptional signature associated with post-treatment AEs; 744 genes were up-regulated.

Our modern PV system infrastructure will connect data traditionally not used in routine safety analysis to help further these types of studies through data enrichment of biobank and genetic data.

C. Process Simplification

Data used in PV has always gone through careful review and analysis. Traditionally, there have been near-equal efforts to test and enrich the data. The promise of ML and related technologies is to reduce the manual efforts required and allow for even more focused effort on specific data activities that will most effectively increase knowledge of the safety characteristics of drugs and vaccines [34].

VI. Enabling Technologies and Emerging Models

A. Process automation of rules-based systems

There are many tasks that are mundane and routine in the process of evaluating safety data. By leveraging our data pathways strategy, one can readily adapt automation of these steps and implement them more quickly than with traditional PV systems infrastructure. While rules-based systems may be considered one of the simpler forms of machine intelligence, there is still much value to be gained from these processes [35].

In March 2020, two new, automated, rules-based processes were released [36]. The first process checks for duplicate reports using predefined sequences, while the second reviews the quality of the data of an incoming safety report by extracting relevant field content and looking for field discrepancies using predefined rules. These methods were shown to significantly reduce the time spent to manually review cases. Over 30,000 safety reports were processed in a single week and it is estimated that the same volume of cases would have required approximately 5,000 person hours to review by hand.

B. Molecular clinical safety intelligence

Safety concerns are common reasons why medicines fail during clinical development. Safety experience with like drugs can inform new drug development and help predict the human safety profile of new drug candidates.

A software tool was developed that warehouses the chemical structures and biological properties of approximately 80,000 compounds to enable molecular clinical safety intelligence. The system enables the analysis of *in vitro*, preclinical, drug metabolism, toxicology, and clinical data to assess the risk of potential toxicity of new candidate drugs [37]. Safety-driven drug design can promote selection of the safest drug candidates for further development.

Tools like these can be accessed to enhance our overall data pathway capabilities. Increasingly, such capability is now obtainable, the outputs of which can then be linked to enhance safety data pathways [38].

C. Data Mining and Machine Learning

In 2011, the FDA Adverse Event Reporting System (FAERS) received more than half a million safety reports [39]. The number of reports filed each year has been growing steadily. In 2021, FAERS recorded over 2.3 million safety reports.

Monitoring of medicines and vaccines is a complex process that cannot be fully automated through rules-based approaches. Much of the activity around processing safety reports is of questionable value in terms of furthering knowledge about patient safety. This motivates the use of more advanced techniques to automate and improve the efficacy of human intervention and manual review of cases. However, in the context of safety report processing, we must be capable of seamlessly supporting both rules-based and ML methods.

The main challenge associated with using ML in safety within this context is developing the safety-specific training data sets needed to "teach" the ML algorithms, that must be dynamic and able to capture changes to the safety environment. The development of training sets is labor-intensive, requiring review of complex safety reports and large amounts of free text to identify and extract relevant variables. The burden of creating better training sets may be alleviated through methods like crowdsourcing.

Additionally, any step in the ML process should clarify how the methods are performing, provide confidence scores, and allow for human intervention when things go wrong [40]. In particular, ML may assist PV-related activities by identifying potential black swan events [41], duplicate case reporting [42], data anomalies or errors, and finding duplicate information in different data sources. Modern PV systems should enable continuous learning from agile data sources, deal with data drift, and allow for course correction when things go wrong.

D. Natural Language processing for automated prioritization of safety literature review

The goal of NLP is for computers to understand the contents of documents. NLP is increasingly being used in PV to glean knowledge from unstructured data (e.g. showing how early identification of acute liver disease from EHRs is possible based on supplementing structure data with NLP extracted concepts from clinical notes [43]).

One hundred percent of the manual review of literature for potential AE cases can be supplanted by NLP to prioritize candidate articles and identify safety reports as shown by Glaser et al. [44]. NLP methods were used to automate and rank literature documents, resulting in a 77% reduction in time in queue for review. All documents identified as relevant were identified in the test dataset, creating tangible gains in efficiency while demonstrating that NLP is can automate the identification of potential AE cases in large volumes of literature.

E. Rapid Query Analysis of Real-Word Data

Rapid query analysis (RQA) methods allow population-based contextualization of outcomes of interest. One example of its use is examining rates of outcomes in an exposed population during specified risk intervals and comparing those rates to those that occur during comparable intervals or in unexposed populations. The results of RQA, however, require further analysis and investigation due to their exploratory nature. RQA can be performed in near real-time, and it may be triggered by emerging internal or external PV-related requests. Various software tools exist to enable RQA of RWD to contextualize observed events [45]. This provides a rapid query capability similar to the FDA's Sentinel network (keeping in mind the limitations of RWD).

This is now a routine capability providing more descriptive and complex analytical analyses across multiple healthcare databases quickly from analysis initiation to results [46].

VII. Multi-modal Safety Monitoring

PV is moving toward a multi-modal model system to leverage the plethora of available data sources and the increasingly sophisticated capabilities of data mining and ML. Multi-modal PV will drive improved data-driven insights into drug and vaccine safety. The data generated by an individual over their lifetime is stored in many platforms and in many forms; and biological & genomic information, medical encounters, diagnoses, procedures, prescriptions and health outcomes may be stored as structured data. While free text associated with these episodes, results of tests and real-time monitoring, and SM posts are stored as unstructured data.

All the building blocks of multi-modal modern PV systems are currently available as individual modules. Linkage across the components is primarily ad hoc and the next phase is to make this more integrated from a user perspective, recognizing that original underlying data will necessarily need to remain fragmented and decentralized.

In summary, the multi-modal systems model provides for increased transparency and efficiency in governance by creating audit trails. It maximizes our ability to interact, analyze and enable data-driven decision making. Finally, the multi-modal system allows for iterative learning from data outputs to inputs.

VIII. Conclusion

By leveraging data pathways, we have described that a next generation, modern PV system benefits from the enhanced linkage of disparate data. The data environment encompasses both traditional and non-traditional data, linking millions of patient lives and data points together to better understand the complexities of patient safety.

This system provides a more holistic approach in both population and personalized patient safety, and enables a higher level of confidence in causality modeling of suspect drug and vaccine event outcomes.

Allowing for more flexibility in the data processing, cleaning and automation of routine safety monitoring processes should foster faster development of advanced safety surveillance methods and encourage more experimentation in the use of advanced data mining and AI/ML methodologies.

Acknowledgment

The authors thank the Modis platform on behalf of GSK for writing assistance provided by Joanne Wolter, graphics support by Gil Costa and manuscript coordination by Carlos Marin.

References

[1] P. Dias, A. Penedones, C. Alves, C. F Ribeiro, and F. B Marques, "The role of disproportionality analysis of pharmacovigilance databases in safety regulatory actions: a systematic review," *Current drug safety*, vol. 10, no. 3, pp. 234–250, 2015.

[2] J. S. Almenoff, E. N. Pattishall, T. Gibbs, W. DuMouchel, S. J. W. Evans, and N. Yuen, "Novel statistical tools for monitoring the safety of marketed drugs," *Clinical Pharmacology & Therapeutics*, vol. 82, no. 2, pp. 157–166, 2007.

[3] A. Bate and S. J. W. Evans, "Quantitative signal detection using spontaneous ADR reporting," *Pharmacoepidemiology and drug safety*, vol. 18, no. 6, pp. 427–436, 2009.

[4] E. G. Brown, L. Wood, and S. Wood, "The medical dictionary for regulatory activities (MedDRA)," *Drug safety*, vol. 20, no. 2, pp. 109–117, 1999.

[5] V. F. Kugener, E. S. Freedland, K. I. Maynard, O. Aimer, P. S. Webster, M. Salas, and M. Gossell-Williams, "Enhancing pharmacovigilance from the US experience: Current practices and future opportunities," *Drug safety*, vol. 44, no. 8, pp. 843–852, 2021.

[6] G. H. Merrill, "The MedDRA paradox," in *AMIA annual symposium proceedings*. American Medical Informatics Association, 2008, pp. 470–474.

[7] A. Bate, E. G. Brown, S. A. Goldman, and M. Hauben, "Terminological challenges in safety surveillance," *Drug safety*, vol. 35, no. 1, pp. 79–84, 2012.

[8] D. Lee and U. Bergman, "The quantification of drug risks in practice," *WHO REGIONAL PUBLICATIONS EUROPEAN SERIES*, pp. 79–79, 1993.

[9] M. D. Solomon and S. R. Majumdar, "Primary non-adherence of medications: lifting the veil on prescription-filling behaviors," pp. 280–281, 2010.

[10] M. Alomar, S. Palaian, and M. M. Al-Tabakha, "Pharmacovigilance in perspective: drug withdrawals, data mining and policy implications," *F1000Research*, vol. 8, 2019.

[11] FDA, "Real-world evidence," https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence, 2018, [Online; accessed 04-May-2022].

[12] G. N. Norén, J. Hopstadius, A. Bate, K. Star, and I. R. Edwards, "Temporal pattern discovery in longitudinal electronic patient records," *Data Mining and Knowledge Discovery*, vol. 20, no. 3, pp. 361–387, 2010.

[13] A. Bate, K. Hornbuckle, J. Juhaeri, S. P. Motsko, and R. F. Reynolds, "Hypothesis-free signal detection in healthcare databases: finding its value for pharmacovigilance," *Therapeutic Advances in Drug Safety*, vol. 10, 2019.

[14] S. J. Reisinger, P. B. Ryan, D. J. O'Hara, G. E. Powell, J. L. Painter, E. N. Pattishall, and J. A. Morris, "Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases," *Journal of the American Medical Informatics Association*, vol. 17, no. 6, pp. 652–662, 2010.

[15] J. U. Stegmann, V. Jusot, O. Menang *et al.*, "Challenges and lessons learned from four years of planning and implementing pharmacovigilance enhancement in sub-Saharan Africa," 2022.

[16] O. Caster, J. Dietrich, M.-L. Kürzinger, M. Lerch, S. Maskell, G. N. Norén, S. Tcherny-Lessenot, B. Vroman, A. Wisniewski, and J. van Stekelenborg, "Assessment of the utility of social media for broad-ranging statistical signal detection in pharmacovigilance: results from the WEB-RADR project," *Drug safety*, vol. 41, no. 12, pp. 1355–1369, 2018.

[17] J. van Stekelenborg, J. Ellenius, S. Maskell, T. Bergvall, O. Caster, N. Dasgupta, J. Dietrich, S. Gama, D. Lewis, V. Newbould *et al.*, "Recommendations for the use of social media in pharmacovigilance: lessons from IMI WEB-RADR," *Drug Safety*, vol. 42, no. 12, pp. 1393–1407, 2019.

[18] G. E. Powell, H. A. Seifert, T. Reblin, P. J. Burstein, J. Blowers, J. A. Menius, J. L. Painter, M. Thomas, C. E. Pierce, H. W. Rodriguez *et al.*, "Social media listening for routine post-marketing safety surveillance," *Drug safety*, vol. 39, no. 5, pp. 443–454, 2016.

[19] Biobanking.com, "10 largest biobanks in the world," https://www.biobanking.com/10-largest-biobanks-in-the-world/, 2021, [Online; accessed 08-Jun-2022].

[20] N. S. Upham, J. H. Poelen, D. Paul, Q. J. Groom, N. B. Simmons, M. P. Vanhove, S. Bertolino, D. M. Reeder, C. Bastos-Silveira, A. Sen *et al.*, "Liberating host–virus knowledge from biological dark data," *The Lancet Planetary Health*, vol. 5, no. 10, pp. e746–e750, 2021.

[21] S. Nogawa, H. Kanamori, K. Tokuda, K. Kawafune, M. Chijiiwa, K. Saito, and S. Takahashi, "Identification of susceptibility loci for adverse events following covid-19 vaccination in the japanese population: A web-based genome-wide association study." *medRxiv*, 2021.

[22] P. R. Epstein, "Climate change and infectious disease: stormy weather ahead?" *Epidemiology*, vol. 13, no. 4, pp. 373–375, 2002.

[23] A. J. Thomas, L. E. Eberly, G. Davey Smith, and J. D. Neaton, "Zip-code-based versus tract-based income measures as long-term risk-adjusted mortality predictors," *American journal of epidemiology*, vol. 164, no. 6, pp. 586–590, 2006.

[24] H. M. Phelos, N. M. Kass, A.-P. Deeb, and J. B. Brown, "Social determinants of health and patient-level mortality prediction after trauma," *Journal of Trauma and Acute Care Surgery*, vol. 92, no. 2, pp. 287–295, 2022.

[25] L. Garcia-Gancedo and A. Bate, "Digital biomarkers for post-licensure safety monitoring," *Drug Discovery Today*, p. 103354, 2022.

[26] R. van de Schoot, J. de Bruin, R. Schram, P. Zahedi, J. de Boer, F. Weijdema, B. Kramer, M. Huijts, M. Hoogerwerf, G. Ferdinands *et al.*, "An open source machine learning framework for efficient and transparent systematic reviews," *Nature Machine Intelligence*, vol. 3, no. 2, pp. 125–133, 2021.

[27] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne *et al.*, "The FAIR guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[28] European Parliament and Council, "DIRECTIVE (EU) 2016/680 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL," https://eur-lex.europa.eu/eli/dir/2016/680/oj/eng, 2021, [Online; accessed 04-May-2022].

[29] V. Mangini, I. Tal, and A.-N. Moldovan, "An empirical study on the impact of GDPR and right to be forgotten-organisations and users perspective," in *Proceedings of the 15th international conference on availability, reliability and security*, 2020, pp. 1–9.

[30] E. Politou, E. Alepis, and C. Patsakis, "Forgetting personal data and revoking consent under the GDPR: Challenges and proposed solutions," *Journal of cybersecurity*, vol. 4, no. 1, p. tyy001, 2018.

[31] M. R. Nelson, H. Tipney, J. L. Painter *et al.*, "The support of human genetic evidence for approved drug indications," *Nature genetics*, vol. 47, no. 8, pp. 856–860, 2015.

[32] M. Kanehisa, M. Araki, S. Goto, M. Hattori *et al.*, "KEGG for linking genomes to life and the environment," *Nucleic acids research*, vol. 36, no. suppl_1, pp. D480–D484, 2007.

[33] B. J. Andersen, B. A. Rosa, J. Kupritz, A. Meite, T. Serge *et al.*, "Systems analysis-based assessment of post-treatment adverse events in lymphatic filariasis," *PLoS neglected tropical diseases*, vol. 13, no. 9, p. e0007697, 2019.

[34] A. Bate and J. U. Stegmann, "Safety of medicines and vaccines–building next generation capability," *Trends in pharmacological sciences*, vol. 42, no. 12, pp. 1051–1063, 2021.

[35] E. Mixson, "Rules based automation explained: What is rules-based automation?" https://www.intelligentautomation.network/intelligent-automation-ia-rpa/articles/rules-based-automation-explained, 2021, [Online; accessed 04-May-2022].

[36] R. Kassekert, M. Easwar, M. Glaser, R. Ventham, and A. Bate, "PNS271 Automation in Routine Use for Data Collection and Processing for Scalable Faster RWE Generation," *Value in Health*, vol. 23, p. S686, 2020.

[37] D. E. Vanderwall, N. Yuen, M. Al-Ansari, J. Bailey, D. Fram, D. V. Green, S. Pickett, G. Vitulli, J. I. Luengo, and J. S. Almenoff, "Molecular clinical safety intelligence: a system for bridging clinically focused safety knowledge to early-stage drug discovery–the GSK experience," *Drug discovery today*, vol. 16, no. 15-16, pp. 646–653, 2011.

[38] T. G. Soldatos, S. Kim, S. Schmidt, L. J. Lesko, and D. B. Jackson, "Advancing drug safety science by integrating molecular knowledge with post-marketing adverse event reports," *CPT: Pharmacometrics & Systems Pharmacology*, vol. 11, no. 5, pp. 540–555, 2022.

[39] S. Weiss-Smith, G. Deshpande, S. Chung, and V. Gogolak, "The FDA drug safety surveillance program: adverse event reporting trends," *Archives of internal medicine*, vol. 171, no. 6, pp. 591–593, 2011.

[40] B. Kompa, J. B. Hakim, A. Palepu, K. G. Kompa, M. Smith, P. A. Bain, S. Woloszynek, J. L. Painter, A. Bate, and A. L. Beam, "Artificial intelligence based on machine learning in pharmacovigilance: a scoping review," *Drug Safety*, vol. 45, no. 5, pp. 477–491, 2022.

[41] O. Kjoersvik and A. Bate, "Black swan events and intelligent automation for routine safety surveillance," *Drug Safety*, vol. 45, no. 5, p. 419–427, 2022.

[42] G. N. Norén, R. Orre, A. Bate, and I. R. Edwards, "Duplicate detection in adverse drug reaction surveillance," *Data Mining and Knowledge Discovery*, vol. 14, no. 3, pp. 305–328, 2007.

[43] L. S. Weiss, X. Zhou, A. M. Walker, A. N. Ananthakrishnan, R. Shen, R. E. Sobel, A. Bate, and R. F. Reynolds, "A case study of the incremental utility for disease identification of natural language processing in electronic medical records," *Pharmaceutical Medicine*, vol. 32, no. 1, pp. 31–37, 2018.

[44] M. Glaser, C. Cranfield, D. Dsouza, A. Duma, K. Hastie, R. Kassekert, and A. Bate, "Automating individual case safety report identification within scientific literature using natural language processing," in *Pharmacoepidemiology And Drug Safety*, vol. 30, 2021, pp. 118–118.

[45] J. Fairburn-Beech, J. Wen, R. Williams, J. Logie, M. Cunnington, and A. Bate, "Rapid cycle analytics to obtain descriptive insight into drug utilisation and background disease rates: building on the sentinel modular programs," in *Pharmacoepidemiology And Drug Safety*, vol. 29, 2020, pp. 616–617.

[46] R. E. Sobel, A. Bate, J. Marshall, K. Haynes, N. Selvam, V. Nair, G. Daniel, J. S. Brown, and R. F. Reynolds, "Do FDA label changes work? Assessment of the 2010 class label change for proton pump inhibitors using the Sentinel System's analytic tools," *Pharmacoepidemiology and Drug Safety*, vol. 27, no. 3, pp. 332–339, 2018.