Big data in clinical research: Present and future

Hyunjoo Kim[1], **Zuo Yen Lee**[2], **Chow Lih Yew**[3]

[1] Clinipace Clinical Research, Seoul,
Republic of Korea
[2] Clinipace Clinical Research, Taipei,
Taiwan (ROC)
[3] Clinipace Clinical Research, Singapore

Correspondence to:

Chow Lih Yew
Clinipace Clinical Research, Singapore
30 Cecil Street #19-08
Prudential Tower, Singapore 049712
+65 68275602
CYew@clinipace.com

Abstract

The clinical research landscape is gradually changing as we enter the era of big data. Big data sources are multiplying as existing sources collide to create expanded platforms that serve wider areas of expertise. Clinical study designs incorporating big data have started to appear and we expect this design phenomenon to grow. Big data offers unprecedented advantages in clinical research, but much remains to be done in assuring accessibility, validity, quality, and privacy protection. For these reasons, medical writers must understand big data, the strengths and the potential limitations of the data used, and should consider big data impact on study design, protocol, and clinical study report authoring. This article provides an overview of big data sources and provides insights on how big data utility could change the clinical-regulatory medical writing landscape.

The changing landscape of clinical research

The overall low generalisability of clinical trial results to routine clinical practice requires new approaches in clinical research.[1] Today, increasing data breadth and depth coupled with advancing data science offer new ways to assess a medicinal product across multiple data sources and at every step of the product's life cycle. We are entering the era of big data. EMA defines big data as "extremely large datasets which may be complex, multi-dimensional, unstructured and heterogeneous, which are accumulating rapidly and which may be analysed computationally to reveal patterns, trends, and associations".[2]

Myriad big data sources are now available, including those considered fit for regulatory decision-making. Table 1 lists example data sources – from the most traditional to relatively newer ones together with their main strengths and limitations.[3–11] This article discusses some of these data sources that are being actively applied in trials.

New ways to use patient registries

Patient registries are "organised systems that use observational methods to collect uniform data over time to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure".[7,12,13] Patient registries could be a powerful tool in clinical studies as we see in VALIDATE-SWEDEHEART (clinicaltrial.gov number: NCT02311231), a prospective study that used the Swedish Coronary Angiography and Angioplasty Registry for both primary data (data collected for a specific, planned study, such as those of randomised clinical trials [RCT]) and secondary data (data already available for another purpose, such as insurance claims data) collection.[14,15] The study used the registry to assess and enrol potential subjects; collect their demographic and baseline data; and randomise subjects to treatment of percutaneous coronary intervention with either bivalirudin or heparin. After treatment, no study visit was required. All study data, including death, myocardial infarction, and major bleeding, were collected directly from the registry, via telephone calls and hospital records.[15,16] This study showcased the advantages of a registry-based RCT in which investigators could enrol many more subjects in a shorter time and the study gained both internal and external validity through the robust design of an RCT that utilised a data source (registry) with higher generalisability than a more traditional design would confer.[17]

Another advantage of working with registries is the accessibility to clinical data for rare diseases, and in which RCTs are often considered unfeasible.[18] Regulators recognise this; in one particular example, due to the low availability of previously untreated haemophilia A patients, the obligation to perform RCTs in these patients has been replaced for marketing authorisation applications of recombinant and human plasma-derived factor VIII products with a new requirement to monitor patients in a registry. The updated guideline also lists the core parameters to support homogeneous data collection across multiple registries – which should be taken into account at an early, pre-authorisation stage of study design. Registries may also be rich sources of secondary data from which suitable data could be extracted to serve as external controls, identify eligible patients, prevent duplicative data

Today, increasing data breadth and depth coupled with advancing data science offer new ways to assess a medicinal product across multiple data sources and at every step of the product's life cycle. We are entering the era of big data.

collection in clinical trials, and provide additional data for benefit/risk assessments.[19]

Using social media for pharmacovigilance

Data from social media are unique because they come directly from patients who have actively decided to share their information.[11] The use of social media in the field of pharmacovigilance and signal detection is not new. One example is the once-hyped Google Flu Trends, which was a web service introduced in 2008 providing estimates of influenza activity by analysing Google search queries. Google developed prediction models that could estimate influenza activity a couple of weeks ahead of the Centres for Disease Control and Prevention's periodic reports.[20] However, the service was terminated in 2015 after algorithmic glitches were detected.[21,22]

Despite Google's failure, numerous studies are testing new ways to utilise comments made on social media to identify potential adverse events; these studies suggest that social media is a promising tool for pharmacovigilance activities, but much work remains to determine its utility and validity.[23,24] Other areas of potential applications include data utilisation in effectiveness assessment, and as a communication tool to gather patient-centric data and to contact patients.[11]

Integrating mHealth in clinical studies

WHO defines mobile health, or mHealth, as the practice of medicine and public health supported by mobile devices for collecting data through symptom monitoring applications, implantable diagnostics, and wearable motion detectors.[25]

The 12-week exploratory Lilly Exploratory Digital Assessment Trial sponsored by Eli Lilly and Apple Inc. was conducted to explore how well mHealth data could discern those with mild cognitive impairment and early Alzheimer's disease from those free of these conditions. The model, applied to the data captured through distributed mobile phones and wearable devices, was able to discern patients from non-patients, suggesting that mild cognitive impairment could be detected in advance.[26] mHealth are in increasing use in clinical studies, acting as data sources for various real-time biometrics and other patient-reported outcomes. Although these novel modelling tools hold tremendous potential, they should be further assessed to ensure that they are "reliable, validated, reproducible, and predictable" to be used for the purpose of regulatory decision-making.[8]

Data collection through mobile devices will likely become more common in clinical studies following the release of the FDA's MyStudies App in 2018 – a digital platform used for multi-site or multi-database studies to collect primary data directly from patients' mobile devices. The application will be linked to an individual's electronic medical record (EMR) and enhanced with additional functions such as e-consent, eligibility test, survey delivery, notifications, and data validation.[27,28] It holds great potential for

Table 1. Sources of big data and their strengths and limitations

Data source	Strengths	Limitations
Clinical trial data (both interventional and non-interventional trials)[3]	● Well-structured data ● High internal quality (integrity/veracity) ● Non-selective sharing ● Publicly accessible trial documents (e.g., protocol, statistical analysis plan)	● Data format and variable definitions across different trials are not standardised
Spontaneous adverse drug reports[4]	● System has a legal/regulatory framework ● Data consistency at a global level ● Competent in detecting new risks of medicines ● Multi-dimensional data; various sources and safety concerns (e.g., medication errors, quality defects, cases of abuse/misuse, occupational exposure)	● Under/over-reporting ● Risk of bias (the safety concern may be the result of increased media attention)
Drug consumption data[5]	● Cover large populations	● Lack individual patient data
Administrative claims data[5,6]	● Data consistency from standardised coding ● Longitudinal record; in EU and in countries with public healthcare service, follow-up period is longer; representative for the source population at a national level ● Provide linkage to data sources ● High quality/complete drug exposure data ● Data on individual's location available for geocoding	● Heterogeneous data in format, variables, quality, and completeness ● Misclassification of diagnosis/exposure/outcome ● Data might not be current ● May lack data on secondary care ● Lack of clinical details ● Data protection legislation may prevent linkage between different health care providers ● Lack of lifestyle/socio-economic factors; lack of control for confounding factors ● Lack over-the-counter drug data
Electronic medical records (EMR)[5,6]	● Diverse clinical data; can complement claims data ● Longitudinal in nature ● Higher validity of diagnosis than claims data from routine use ● Provide linkage to data sources	● Heterogeneous data in format, variables, quality, and completeness ● Patient privacy concerns ● May contain only one type of care setting (primary or secondary) ● Lack of lifestyle/socio-economic factors; lack of control for confounding factors ● Lack over-the-counter drug data
Patient registry [5–8]	● Data consistency ● Established, large registry programmes ● Able to observe the course of disease and effects of new treatments	● Limited to specific procedures, diseases, or settings ● Data might not be current ● Discrepancy between collected data and data requested by the regulatory authority ● Inconsistent data and varying quality across registries ● May need source data verification
Biomarkers (including any "omics" data)[8–10]	● Precision medicine ● Identification of unique molecular markers of disease/responsiveness to medications	● High genetic variation ● False positives/negatives ● Further validation needed to associate biomarker data to patient outcomes ● Lack of publicly accessible, clinically meaningful information from the genomic database ● Lack of data standardisation ● Patient privacy concerns, especially for patients with rare diseases ● Heterogeneous data

Data source	**Strengths**	**Limitations**
Medical imaging[3,6]	● General data consistency ● Widely used in clinical trials; unexplored potential in various therapeutic areas	● Lack of accessibility ● Ethical issues related to data sharing ● Challenges on analysing/integrating imaging data with other data sources
Social media[6,11]	● Wide reach of the internet ● Various types of data ● Result of active sharing from patients	● Heterogeneous data ● Lack of specificity in general social media; data prone to bias ● Limited follow up; difficult to verify/validate ● Lacks consideration of the characteristics of the patients included ● Lacks Good Clinical Practice adherence ● Lacks validity and reliability ● Patient privacy concerns
Mobile health (mHealth) and wearable devices[6,8,11]	● Collected biometrics data may allow control for confounding factors ● Patient-centric data ● Continuous data from real life (vs. episodic data restricted to healthcare setting) ● Data readily available for research purposes; platforms support central data management, analysis, and reporting and can often be directly linked to an electronic case report form ● Devices can monitor parameters to calculate/monitor drug dose	● Further validation needed to discern clinically important 'signals'; unknown sensitivity of the collected data ● Precision does not necessarily mean accuracy ● Output is highly variable across different types of device ● Output depends on the level of user interaction ● Lack of familiarity with interpretation of the data ● Potential challenges in timing of surveys in relation to other healthcare data ● Patient privacy and security concerns, e.g., hacking

pragmatic trials – which are evidential for the use of a clinical practice intervention and may, therefore, guide policy-making, and observational trials – trials without an intervention.[29]

In the wake of the recent coronavirus disease 2019 (COVID-19) pandemic, many mHealth initiatives have been developed for gathering information to help manage the outbreak. For example, Scripps Research Translational Institute launched the DETECT study in March 2020 to collect health data through wearables like smartwatches and activity trackers for a public health surveillance programme for early detection of viral diseases; at the same time, Stanford Medicine also initiated the COVID-19 Wearable Study that serves the same purpose.[30–33] In April 2020, the two platforms joined forces, together with Fitbit, to create a consortium which will aggregate data for knowledge sharing and drive wearables research.[32,34]

Larger data platforms

Existing data are expanding, and are also being linked across various networks, creating larger data platforms. Sentinel is FDA's national safety surveillance system to monitor its regulated medical products. The system extracts electronic health records (EHRs) from various networks, mostly from health insurers.[35] Sentinel is now collaborating with over 40 other networks across three centres – Sentinel Operations Centre, Innovation Centre, and Community Building and Outreach Centre – to cover wider areas of scientific expertise, improve technologies translation, and encourage communication and collaboration.[35]

A Sentinel collaborator, Patient-Centered Outcomes Research Network (PCORnet®), is a partnership of over 10 networks. PCORnet® contains more EMR data with various types of individual patient data, including laboratory test results, vital signs, biospecimen data, genomic data, and patient satisfaction data.[36,37] ADAPTABLE (clinicaltrials.gov number: NCT02697916), a pragmatic clinical study that

compares the effectiveness of two doses of aspirin (81 mg and 325 mg) in approximately 20,000 patients, uses the existing EHRs and a web-based patient portal in PCORnet® to identify eligible patients, obtain consents, randomise, and follow up with patients.[38]

No such platform is available in the EU yet. However, recently the Heads of Medicines Agencies (HMA)/EMA Task Force on Big Data proposed their plans to establish an EU platform, namely Data Analysis and Real-World Interrogation Network (DARWIN), to access and analyse healthcare data from across the EU to inform regulatory decision-making. This initiative is one of the many efforts undertaken by the EMA to optimise the use of big data in medicines regulation.[39] During the recent COVID-19 pandemic, international regulators and experts from WHO and European Commission acknowledged the value of real-world data from COVID-19 observational studies and how these data could complement clinical trials in finding solutions to prevent and treat COVID-19. Public platforms, such as EU PAS Register and ClinicalTrials.gov, were identified as suitable platforms to share and exchange information about COVID-19 observational studies.[40,41]

On the horizon

In the future, it may be possible to create a complete, longitudinal record of an individual starting from the omics level. Collaborations between academia, companies, and regulatory authorities nationally and internationally culminated in the initiative of the Electronic Medical Records and Genomics (eMERGE) Network. Since 2007, the Network has brought together researchers in genomics, statistics, ethics, informatics, and clinical medicine areas with the goal to combine a biological materials repository with EMR systems for research at the genomic level.[9,42]

As regulatory medical writers, we need to consider how best to leverage big data into our work outputs.

About 90% of medical data are in the form of images captured with increasingly higher quality and improved resolution. Much of these voluminous data are stored unanalysed.[43] To utilise these data, the UK Biobank Imaging Study aims to develop longitudinal records from volunteers consisting of their brain, heart, and body imaging data; biomarker and genetic analysis results; physical measurements; and self-reported health and lifestyle data. These records can also be linked to the individual's National Health Service records.[44,45] In April 2020, the UK Biobank announced that it would grant access to the health data of its 500,000 participants to researchers for health-related research. These data include results of COVID-19 tests, primary care data, hospital episodes, and intensive care data.[46]

The exponential advances in personal omics profiling, coupled with the increasing amount of high-frequency data using wearable devices, omics data, imaging data, as well as enlarging platforms and dynamic patient-centred interfaces are set to greatly affect how we conduct clinical research.

Guidelines for using big data in regulatory decision-making

The message from regulators is that we must embrace the use of big data. In January 2017, the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use Assembly (ICH) endorsed the ICH reflection paper entitled "ICH Reflection on "GCP Renovation": Modernisation of ICH E8 and Subsequent Renovation of ICH E6"[47] to address the increasing diversity of clinical trial designs and data sources being employed.

FDA has also published their final guidance on the "Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices" (August 2017),[48] "Use of Electronic Health Record Data in Clinical Investigations" (July 2018),[49] and a draft guidance for "Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drugs and Biologics" (May 2019)[50] to guide and encourage the use of big data in the industry.

Using big data in medical writing

Using big data to improve study efficiency

As regulatory medical writers, we need to consider how best to leverage big data into our work outputs. Big data could be applied at any stage of study design, through to enrolment and data analysis.[51] Before study initiation, we could use existing big data, sourced from registries and EHRs to help identify an appropriate target population, define more targeted eligibility criteria, determine if a sufficient number of subjects are likely to be available, and even obtain baseline data directly from the data sources. Thus, study efficiency could be greatly improved by reducing the cost of and time for recruitment, reducing patient attrition and minimising possible changes to the protocol down the line.[51,52]

Heterogeneity in the data is the intrinsic underlying issue in most data sources and this aspect should be thought through in the study design and statistical analysis, in consultation with biostatistical colleagues – our natural partners in analysis and reporting.

Throughout the study, data collected from various digital sources (such as mobile applications and wearable devices) may be available faster than data collected by traditional methods, thereby allowing for prompt futility analyses in a study or benefit/risk assessment in a post-marketing surveillance, hence more rapid decision-making. Big data platforms like registries also help track patients during study follow-up under their usual care routines, thus minimising patients being lost to follow-up and reducing missing data during a study.[51]

Using big data in study design

For some diseases where patient enrolment may be problematic (e.g., rare diseases) or randomising patients to the control group may be unethical (e.g., cancer), using an external control group can be considered. An external control group refers to subjects who are selected from an external source, e.g., existing clinical trial data and EHRs. The biggest challenge of using an external control is bias control. FDA suggests the use of external controls only under certain conditions, e.g., when we expect distinct treatment effects between the test and external control groups. External control should be selected from data sources that are most appropriate to the study purpose and should align, as much as possible, with the study eligibility criteria to minimise potential confounding and selection biases.[53] Another important consideration is the availability of similar endpoint assessments between the test group and the external control group to allow comparison between them. In this case, external control groups derived from existing clinical studies with similar purposes may be more applicable than those from EHRs or registries.[54]

Heterogeneity in the data is the intrinsic underlying issue in most data sources and this aspect should be thought through in the study design and statistical analysis, in consultation with biostatistical colleagues – our natural partners in analysis and reporting. When selecting which databases to use, accessibility, storage, and quality of the data are paramount considerations as they ensure reliability and validity of the data. We must be mindful while extracting data that they may contain missing information that could bias our interpretation of the data. For example, missing data does not mean the absence of an event; the absence of smoking status in the medical record may not mean the patient is not a smoker.

Designing a study using big data requires rather different elements and methods from that of traditional RCTs. Existing guidelines such as the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance "Guide on Methodological Standards in Pharmacoepidemiology",[14] its protocol checklist,[55] and the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist[56] provide some starting guidance for medical writers on important elements that should be considered, such as procedures for selecting target population, defining covariates, methods to address each type of bias, and related statistical analyses.

It is important to remember that study design is always a collaborative endeavour with colleagues in other functional areas such as biostatistics and medical affairs. As medical writers, we can and should be influencers. We can raise awareness of the potential of big data in study design to ensure that all stakeholders consider its practical utility.

Permission from a subject to use his or her personal health data, in the form of informed consent or authorisation with pre-defined purposes, is required before data collection.

Permission for secondary use of personal data

Permission from a subject to use his or her personal health data, in the form of informed consent or authorisation with pre-defined purposes, is required before data collection. Big data analytics seek patterns and associations from big datasets that are often generated by pooling or linking data from various studies and databases. Therefore, secondary use, i.e. use of existing data collected for other purposes, is more common for big data analytics.

Personal data that will be collected, processed, or stored within the EU need to comply with the General Data Protection Regulation (GDPR). Under the GDPR, new consent is not required for the processing and secondary use of personal data for scientific research purposes provided specific adequate safeguards and conditions are adhered to, such as pseudonymisation.[57,58] GDPR also acknowledges that it is "often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection" and allows subjects to consent to a more general purpose. Nonetheless, in addition to the specific consent, GDPR requires that a separate consent with the general areas of secondary research be specified and options to "consent only to certain areas of research or parts of the research projects" be provided before data collection.[59,60] Of note, the use of de-identified personal data does not fall within the scope of the GDPR.[61]

In the US, the "Revised Common Rule" that took effect in January 21, 2019, accelerates the secondary use of data through the introduction of Broad Consent. Broad Consent allows subjects to consent to unspecified future research that may store, maintain, or use their identifiable private information or identifiable biospecimens for secondary research before data collection. Important information, such as the types of research that may be conducted, information that may be used, the institutions that may reuse the information, and the time frame of the consent must be included in the Broad Consent.[62,63]

Under the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, "covered entities", including health plans, health care clearinghouses, and health care providers, should obtain an individual's written authorisation for any use of protected health information (PHI) for secondary research.[64,65] Core elements, such as the purpose of the use, the specific information to be used, the persons who can use the PHI, and the time frame of the authorisation must be included in the authorisation.[65] An Institutional Review Board or privacy board waiver of authorisation is required to use PHI for research purposes if individual authorisation is not available.[66] Currently, there are no restrictions (i.e. neither consent nor HIPAA authorisation is required) on the use of de-identified health information.[67,68]

> Big data is expected to offer unprecedented advantages in every step of clinical research by providing alternative study design, improving study efficiency, and accelerating regulatory decision-making.

Conclusion

Big data is expected to offer unprecedented advantages in every step of clinical research by providing alternative study design, improving study efficiency, and accelerating regulatory decision-making. At the same time, they also pose new challenges, especially in ensuring data quality and privacy protection. An enormous amount of health data has become available during the recent COVID-19 pandemic, and we have directly experienced how researchers and regulators across the world use big data in the fight against COVID-19. Undoubtedly, we as medical writers should start honing the necessary skills and competencies to better prepare ourselves as we embrace the era of big data.

Acknowledgements

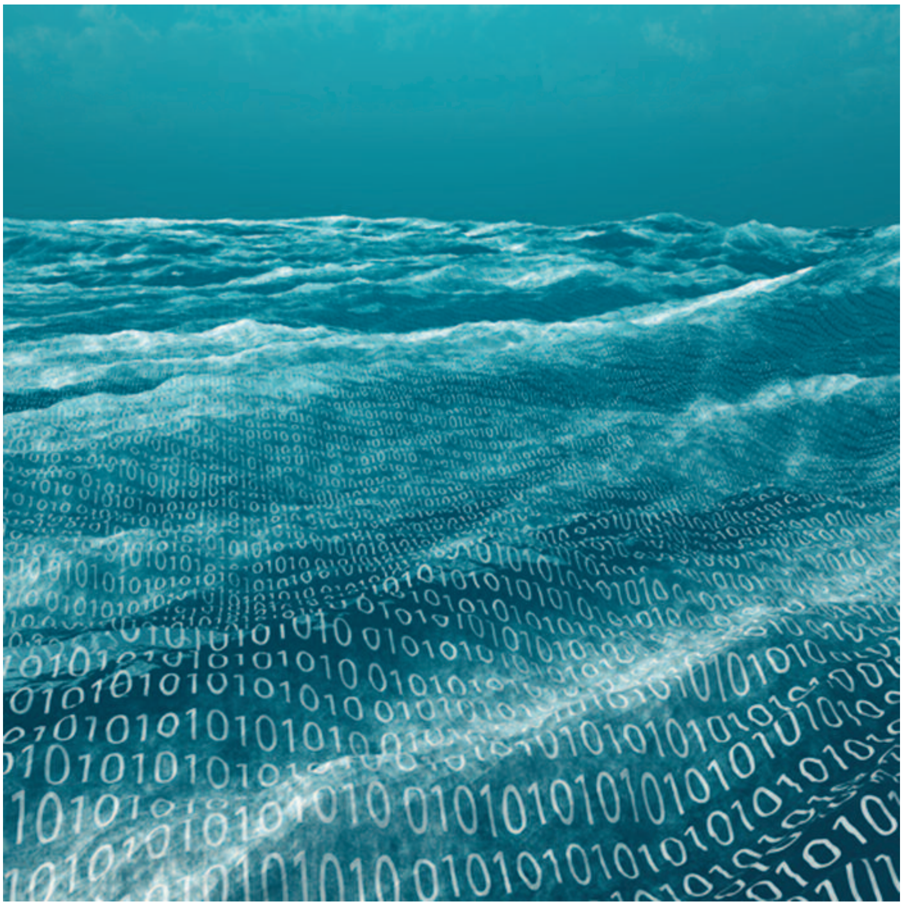
The authors would like to thank Samina Hamilton for her review and input in the draft manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Bonell C, Oakley A, Hargreaves J, Strange V, Rees R. Assessment of generalisability in trials of health interventions: suggested framework and systematic review. BMJ. 2006;333(7563):346–9.
2. European Medicines Agency. Big data. 1995-2020 [cited 2020 Feb 26]. Available from: https://www.ema.europa.eu/en/about-us/how-we-work/big-data
3. Kiviniemi V, Rosso A, Szabone ZC, Nyeland M, Fuglerud P. Clinical Trial and Imaging Subgroup report [Internet]. [cited 2020 Feb 26]. Available from: https://www.ema.europa.eu/en/documents/report/clinical-trial-imaging-subgroup-report_en.pdf
4. García CH, Pinheiro L, Maciá MÁ, et al. Spontaneous Adverse Drug Reactions Subgroup report [Internet]. [cited 2020 Feb 26]. Available from: https://www.ema.europa.eu/en/documents/report/spontaneous-adverse-drug-reactions-subgroup-report_en.pdf
5. Rosso A, Pacurariu A, Cave A, et al. Observational data (Real World Data) Subgroup report [Internet]. [cited 2020 Feb 26]. Available from https://www.ema.europa.eu/en/documents/report/observational-data-real-world-data-subgroup-report_en.pdf
6. Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. Nat Rev Cardiol. 2016;13(6):350–9.
7. Gliklich RE, Leavy MB, Dreyer NA, editors. Registries for Evaluating Patient Outcomes: A User's Guide. (Prepared by the Outcome DEcIDE Center [Outcome Sciences, Inc. dba Outcome] under Contract No. HHSA29020050035I TO1.) AHRQ Publication No. 07-EHC001-1 [Internet]. Rockville, MD: Agency for Healthcare Research and Quality; 2007 [cited 2020 Feb 26]. Available from: https://effectivehealthcare.ahrq.gov/sites/default/files/pdf/registries-guide_research.pdf
8. emainfo. Welcome, and update on work of the HMA/EMA Joint Big Data Task Force on Big Data [video file]. 2018 [cited 2020 Feb 26]. Available from: https://www.youtube.com/watch?list=PL7K5dNgKnawbkBU1mXLDWAnfkFOzSeDRo&v=mUDdJH0lf2I&feature=emb_logo
9. Meulendijks D, Deforce D, Ovelgönne H, König R, Pasmooij M. Genomics Genetics, Transcriptomics and Epigenetics Subgroup report [Internet]. [cited 2020 Feb 26]. Available from https://www.ema.europa.eu/en/ documents/report/genomics-genetics-transcriptomics-epigenetics-subgroup-report_en.pdf
10. König R, Cave A, Goldammer M, Meulendijks D. Bioanalytical Omics Subgroup report [Internet]. [cited 2020 Feb 26]. Available from: https://www.ema.europa.eu/en/documents/report/bioanalytical-omics-subgroup-report_en.pdf
11. Donegan K, Ovelgonne H, Flores G, Fuglerud P, Georgescu A. Social Media and M-Health Data Subgroup report [Internet]. [cited 2020 Feb 26]. Available from: https://www.ema.europa.eu/en/documents/report/social-media-m-health-data-subgroup-report_en.pdf
12. Zaletel M, Kralj M, editors. Methodological Guidelines and Recommendations for Efficient and Rational Governance of Patient Registries [Internet]. Ljubljana: National Institute of Public Health; 2015 [cited 2020 Feb 26]. Available from: https://ec.europa.eu/health/sites/health/files/ehealth/docs/patient_registries_guidelines_en.pdf
13. Polygenis D, Frame D, Blanchette C, editors. ISPOR Taxonomy of Patient Registries: Classification, Characteristics and Terms. Lawrenceville, NJ: ISPOR; 2013.
14. The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance. Guide on Methodological Standards in Pharmacoepidemiology (Revision 7). Reference No.: EMA/95098/2010 [Internet]. [cited 2020 Feb 26]. Available from: http://www.encepp.eu/standards_and_guidances/documents/ENCePPGuideonMethStandardsinPE_Rev7.pdf
15. ClinicalTrials.gov. Bivalirudin vs Heparin in NSTEMI and STEMI in Patients on Modern Antiplatelet Therapy in SWEDEHEART (VALIDATE). 2017 [cited 2020 Feb 26]. Available from: https://clinicaltrials.gov/ct2/show/NCT02311231
16. Erlinge D, Omerovic E, Fröbert O, et al. Bivalirudin versus Heparin Monotherapy in Myocardial Infarction. N Engl J Med. 2017;377(12):1132–42.
17. Lauer MS, D'Agostino RB Sr. The Randomized Registry Trial – The Next Disruptive Technology in Clinical Research?. N Engl J Med. 2013;369(17):1579–81.
18. McGettigan P, Alonso Olmo C, Plueschke



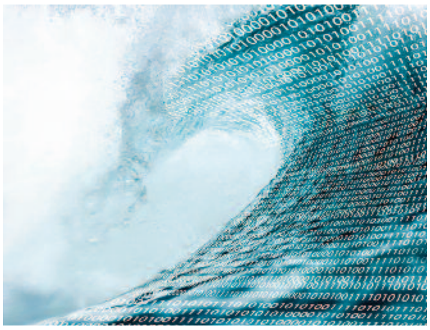
K, et al. Patient Registries: An Underused Resource for Medicines Evaluation: Operational proposals for increasing the use of patient registries in regulatory assessments. Drug Saf. 2019;42(11): 1343–51.

19. European Medicines Agency. Guideline on the clinical investigation of recombinant and human plasma-derived factor VIII products (Revision 2). Reference No.: EMA/CHMP/BPWP/144533/2009 [Internet]. [cited 2020 Feb 26]. Available from: https://www.ema.europa.eu/en/clinical-investigation-recombinant-human-plasma-derived-factor-viii-products
20. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457(7232):1012–4.
21. Cook S, Conrad C, Fowlkes AL, Mohebbi MH. Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. PLoS One. 2011;6(8):e23610.
22. Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. PLoS Comput Biol. 2013;9(10):e1003256.
23. Anderson LS, Bell HG, Gilbert M, et al. Using Social Listening Data to Monitor Misuse and Nonmedical Use of Bupropion: A Content Analysis. JMIR Public Health Surveill. 2017;3(1):e6.
24. Powell GE, Seifert HA, Reblin T, et al. Social Media Listening for Routine Post-Marketing Safety Surveillance. Drug Saf. 2016;39(5):443–54.
25. World Health Organization. mHealth: New horizons for health through mobile technologies (Global Observatory for eHealth series - Volume 3) [Internet]. 2011 [cited 2020 Feb 26]. Available from https://www.who.int/goe/publications/ehealth_series_vol3/en/
26. Chen R, Jankovic F, Marinsek N, et al. Developing Measures of Cognitive Impairment in the Real World from Consumer-Grade Multimodal Sensor Streams. In: The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19); 2019 August 4-8; Anchorage, AK, US. New York: ACM; 2019. p2145-2155. Available from https://evidation.com/wp-content/uploads/2019/08/developing-measures-of-cognitive-impairment-in-the-real-world-from-consumer-grade-multimodal-sensor-streams.pdf
27. U.S. Food & Drug Administration. FDA's MyStudies Application (App). 2019 [cited 2020 Feb 26]. Available from: https://www.fda.gov/drugs/science-and-research-drugs/fdas-mystudies-application-app
28. U.S. Food & Drug Admnistration. Mobile App Quick Overview for Research [Internet]. 2019 [cited 2020 Feb 26]. Available from: https://www.fda.gov/media/119834/download.
29. Ford I, Norrie J. Pragmatic Trials. N Engl J Med. 2016;375(5):454–63.
30. Scripps Research Translational Institute. Scripps Research invites public to join app-based DETECT study, leveraging wearable data to potentially flag onset of viral illnesses. [Internet]. 2020 [cited 2020 Apr 19]. Available from: https://www.scripps.edu/news-and-events/press-room/2020/20200325-detect-study-viral-illnesses.html
31. Scripps Research Translational Institute. DETECT Health Study. 2020 [cited 2020 Apr 19]. Available from: https://detectstudy.org/
32. Stanford Medicine. Stanford Medicine scientists hope to use data from wearable devices to predict illness, including COVID-19. [Internet] 2020 [cited 2020 Apr 19]. Available from https://med.stanford.edu/news/all-news/2020/04/wearable-devices-for-predicting-illness-.html
33. Stanford Medicine. Fight COVID-19 through the power of people. 2020 [cited 2020 Apr 19]. Available from: https://innovations.stanford.edu/#section-wearables
34. mHealth Intelligence. SRTI, Stanford Health Launch mHealth Consortium for COVID-19 Research. [Internet] 2020 [cited 2020 Apr 19]. Available from: https://mhealthintelligence.com/news/srti-stanford-health-launch-mhealth-consortium-for-covid-19-research
35. U.S. Food & Drug Admnistration. FDA's Sentinel Initiative. 2019 [cited 2020 Feb

26]. Available from https://www.fda.gov/safety/fdas-sentinel-initiative
36. PCORnet. PCORnet Common Data Model. 2019 [cited 2020 Feb 26]. Available from: https://pcornet.org/data-driven-common-model
37. PCORnet. Real World Data for Clinical Research: A PCORnet Workshop with the Pharmaceutical and Biologics Industry [Internet]. Washington, DC, US: PCORI; 2019 [cited 2020 Feb 26]. Available from: https://www.pcori.org/events/2015/real-world-data-clinical-research-pcornet-workshop-pharmaceutical-and-biologics-industry
38. Johnston A, Jones WS, Hernandez AF. The ADAPTABLE Trial and Aspirin Dosing in Secondary Prevention for Patients with Coronary Artery Disease. Curr Cardiol Rep. 2016;18(8):81.
39. European Medicines Agency. Ten recommendations to unlock the potential of big data for public health in the EU. 2020 [cited 2020 Feb 26]. Available from: https://www.ema.europa.eu/en/news/ten-recommendations-unlock-potential-big-data-public-health-eu
40. European Medicines Agency. Global regulators discuss observational studies of real world data for COVID-19 medicines. [Internet] 2020 [cited 2020 Apr 19]. Available from: https://www.ema.europa.eu/en/news/global-regulators-discuss-observational-studies-real-world-data-covid-19-medicines
41. European Medicines Agency. Meeting highlights from ICMRA global regulatory workshop on COVID-19 observational studies and real world data. [Internet] 2020 [cited 2020 Apr 19]. Available from: https://www.ema.europa.eu/en/news/meeting-highlights-icmra-global-regulatory-workshop-covid-19-observational-studies-real-world-data
42. National Human Genome Research Institute. Electronic Medical Records and Genomics (eMERGE) Network. 2020 [cited 2020 Feb 26]. Available from https://www.genome.gov/Funded-Programs-Projects/Electronic-Medical-Records-and-Genomics-Network-eMERGE
43. GE Healthcare. Beyond Imaging: the Paradox of AI and Medical Imaging Innovation. 2018 [cited 2020 Feb 26]. Available from: https://www.gehealthcare.com/article/beyond-imagingthe-paradox-of-ai-and-medical-imaging-innovation
44. Nature. UK Biobank data on 500,000 people paves way to precision medicine. 2018 [cited 2020 Feb 26]. Available from: https://www.nature.com/articles/d41586-018-06950-9
45. UK Biobank. About UK Biobank. 2019 [cited 2020 Feb 26]. Available from: http://www.ukbiobank.ac.uk/about-biobank-uk/
46. UK Biobank. UK Biobank makes infection and health data available to tackle COVID-19. [Internet] 2020 [cited 2020 Apr 19]. Available from: https://www.ukbiobank.ac.uk/2020/04/covid/
47. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. ICH Reflection on "GCP Renovation": Modernization of ICH E8 and Subsequent Renovation of ICH E6. 2017 [cited 2020 Feb 26]. Available from: https://admin.ich.org/sites/default/files/2019-04/ICH_Reflection_paper_GCP_Renovation_Jan_2017_Final.pdf
48. U.S. Food & Drug Administration. Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices. Reference No.: FDA-2016-D-2153 [Internet]. 2017 [cited 2020 Feb 26]. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-real-world-evidence-support-regulatory-decision-making-medical-devices
49. U.S. Food & Drug Admnistration. Use of Electronic Health Record Data in Clinical Investigations. Reference No.: FDA-2016-D-1224 [Internet]. 2018 [cited 2020 Feb 26]. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-electronic-health-record-data-clinical-investigations-guidance-industry
50. U.S. Food & Drug Administration. Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drugs and Biologics. Reference No.: 2019-09529 [Internet]. 2019 [cited 2020 Feb 26]. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/submitting-documents-using-real-world-data-and-real-world-evidence-fda-drugs-and-biologics-guidance
51. Mc Cord KA, Al-Shahi Salman R, Treweek S, et al. Routinely collected data for randomized trials: promises, barriers, and implications. Trials. 2018;19(1):29.
52. Gilder JR, Big Data in Healthcare: The next frontier in clinical research, study trial recruitment, and quality assessment. The Monitor. 2012;26:55–8.
53. U.S. Food & Drug Administration. E10 Choice of Control Group and Related Issues in Clinical Trials. Reference No.: FDA-2013-S-0610 [Internet]. 2001 [cited 2020 Feb 26]. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e10-choice-control-group-and-related-issues-clinical-trials
54. Beckers F, Capra W, Cassidy A, et al. Characterizing the use of external controls for augmenting randomized control arms and confirming benefit. Proceedings of the Friends 2019 Annual Meeting Panel 1; 2019 Nov 12; Washington, DC, US. Washington, DC: Friends of Cancer Research. 2019 [cited 2020 Feb 26]. Available from https://www.focr.org/sites/default/files/pdf/Panel-1_External_Control_Arms2019AM.pdf
55. The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance. Checklist for Study Protocols (Revision 4). Reference No.: EMA/540136/2009 [Internet]. 2018 [cited 2020 Feb 26]. Available from: http://www.encepp.eu/standards_and_guidances/checkListProtocols.shtml
56. STROBE Statement. STROBE checklists. 2007 [cited 2020 Feb 26]. Available from https://www.strobe-statement.org/index.php?id=available-checklists
57. Principles relating to processing of personal data. Regulation (EU) 2016/679 of the European Parliament and of the Council of

27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, OJ L 119, 4.5.2016, Article 5(1)(b).

58. Safeguards and derogations relating to processing for scientific research. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, OJ L 119, 4.5.2016, Article 89(1).
59. General requirement for consent to general purpose. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, OJ L 119, 4.5.2016, Recital 33.
60. European Commission Directorate-General for Health and Food Safety. Question and Answers on the interplay between the Clinical Trials Regulation and the General Data Protection Regulation [Internet]. 2019 [cited 2020 Feb 26]. Available from: https://ec.europa.eu/health/sites/health/files/files/documents/qa_clinicaltrials_gdpr_en.pdf
61. Use of de-identified personal data. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, OJ L 119, 4.5.2016, Recital 26.
62. General requirements for broad consent. Revised Common Rule. 45 Code of Federal Regulation 46.116. [cited 2020 Feb 29]. Available from: https://www.ecfr.gov/cgi-bin/text-idx?SID=1e1c2b5b18e1f4cc624695c38e6057f2&mc=true&node=pt45.1.46&rgn=div5#se45.1.46_1116
63. Office for Human Research Protections. Revised Common Rule Q&As [Internet]. 2018 [cited 2020 Feb 26]. Available from: https://www.hhs.gov/ohrp/education-and-outreach/revised-common-rule/revised-common-rule-q-and-a/index.html
64. Definition of 'covered entities'. HIPAA Privacy Rule. 45 Code of Federal Regulation 160.103. 2019 [cited 2020 Feb 29]. Available from: https://www.ecfr.gov/cgi-bin/text-idx?SID=1e1c2b5b18e1f4cc624695c38e6057f2&mc=true&node=pt45.2.160&rgn=div5#se45.2.160_1103
65. General requirements for authorization. HIPAA Privacy Rule. 45 Code of Federal Regulation 164.508. [cited 2020 Feb 29]. Available from: https://www.ecfr.gov/cgi-bin/text-idx?SID=b35279ee7bdca503d9385c7e6c09b078&mc=true&node=pt45.2.164&rgn=div5#se45.2.164_1508
66. Waiver of authorization. HIPAA Privacy Rule. 45 Code of Federal Regulation 164.512(i)(1)(i). [cited 2020 Feb 29]. Available from: https://www.ecfr.gov/cgi-bin/text-idx?SID=b35279ee7bdca503d9385c7e6c09b078&mc=true&node=pt45.2.164&rgn=div5#se45.2.164_1512
67. Use of de-identified information. Revised Common Rule. 45 Code of Federal Regulation 46.104(d)(4)(ii). [cited 2020 Feb 29]. Available from: https://www.ecfr.gov/cgi-bin/text-idx?SID=1e1c2b5b18e1f4cc624695c38e6057f2&mc=true&node=pt45.1.46&rgn=div5#se45.1.46_1104
68. Use of de-identified information. HIPAA Privacy Rule. 45 Code of Federal Regulation 164.502(d)(2). [cited 2020 Feb 29]. Available from: https://www.ecfr.gov/cgi-bin/text-idx?SID=b35279ee7bdca503d9385c7e6c09b078&mc=true&node=pt45.2.164&rgn=div5#se45.2.164_1502



Author information

Hyunjoo (Mary) Kim is a pharmacist-turned-medical writer who has a great interest in big data. After graduating with an MSc in molecular pharmacology, she worked at Pfizer Korea as an outcomes research and real-world data specialist. In December 2017, she started her career as a medical and regulatory writer at Clinipace Korea.

Zuo Yen Lee graduated with a PhD in biology from ETH Zurich. She found passion in medical writing after a decade of training in academic research and the veterinary diagnostic industry. She is a medical and regulatory writer at Clinipace Taiwan since January 2017. She enjoys every bit of it and is always looking to broaden her horizons.

Chow Lih Yew has been a medical and regulatory writer at Clinipace Singapore since 2018. She enjoys following the latest trends in clinical research and drug development. Before this, she gained over 10 years of experience in academic research in Japan and Switzerland, where she earned her PhD in molecular biology.