Injury xxx (xxxx) xxx

Contents lists available at ScienceDirect

Injury

journal homepage: www.elsevier.com/locate/injury

Registries: Big data, bigger problems?

Luc Rubinger[a,1,*], Seper Ekhtiari[a,b,1], Aaron Gazendam[a,b,1], Mohit Bhandari[a,b,1]

[a] Division of Orthopaedics, Department of Surgery, McMaster University, Hamilton, ON Canada
[b] Centre for Evidence-Based Orthopaedics, 293 Wellington St. N, Suite 110, Hamilton, ON L8L 8E7 Canada

ARTICLE INFO

Article history:
Accepted 4 December 2021
Available online xxx

Keywords:
Registry
Big-data
registry-based RCT

ABSTRACT

Patient registries have grown in size and number along with general computing power and digitization of the healthcare world. In contrast to databases, registries are typically patient data systematically created and collected for the express purpose of answering health-related questions. Registries can be disease-, procedure-, pathology-, or product-based in nature. Registry-based studies typically fit into Level II or III in the hierarchy of evidence-based medicine. However, a recent advent in the use of registry data has been the development and execution of registry-based trials, such as the TASTE trial, which may elevate registry-based studies into the realm of Level I evidence. Some strengths of registries include the sheer volume of data, the inclusion of a diverse set of participants, and their ability to be linked to other registries and databases. Limitations of registries include variable quality of the collected data, and a lack of active follow-up (which may underestimate rates of adverse events). As with any study type, the intended design does not automatically lead to a study of a certain quality. While no specific tool exists for assessing the quality of a registry-based study, some important considerations include ensuring the registry is appropriate for the question being asked, whether the patient population is representative, the presence of an appropriate comparison group, and the validity and generalizability of the registry in question. The future of clinical registries remains to be seen, but the incorporation of big data and machine learning algorithms will certainly play an important role.

Introduction – what is a registry?

Patient registries, as the health community uses them, are data systems organized in a way that allows the prospective collection and utilization of observational and clinical data to flexibly assess specific outcomes for a population with a stated scientific, clinical or policy purpose. These populations are usually defined by a particular condition, disease, or exposure [1,2]. Meanwhile, data for health-related registries are sourced from: patient-reported data, physician-reported data, medical chart abstraction, electronic medical records (EMRs), and administrative and organization databases, among others. The distinction must be drawn between databases and registries; a database is merely an electronic set of data that is neither systematized nor organized for the explicit use for answering health-related questions. The development and innovation of electronic data collection systems, with the seemingly exponential rise in data points being created along with the staggering size of databases, has led to a natural increase in the number of registries developed and used for research, policy, and administrative purposes. But with this staggering rise in registry utilization, it is important for clinicians, and all users of registry-based literature, to understand the strengths, limitations, and future directions of this important information source, along with how to assess the quality registry data.

Registry classification

According to Gliklich et al., [1] classification of registries can occur based on the methods used to collect the included data within the registry. These include disease-, procedure- or pathology-specific registries, administrative-, health-systems-based or combined/linked registries, or product registries. Examples of disease-specific registry, where data is gathered on a defined cohort of patients, include the well-developed national cancer registries of Sweden, Denmark, Finland, Iceland, and Norway, that prospectively collect data on cancer diagnoses, treatment, and outcomes. Taking this example one step further, the NORDCAN-program presents

* Corresponding author at: Center for Evidence-Based Orthopaedics, Division of Orthopaedics, Department of Surgery, McMaster University, 293 Wellington Street North, Suite 110 Hamilton, ON L8L 8E7, Canada.
E-mail address: luc.rubinger@medportal.ca (L. Rubinger).

[1] All authors have made substantial contributions to all of the following: (1) the conception and design of the paper, (2) drafting the article or revising it critically for important intellectual content, (3) final approval of the version to be submitted. The manuscript, including related data, figures and tables has not been previously published and is not under consideration elsewhere.

https://doi.org/10.1016/j.injury.2021.12.016
0020-1383/

projections of cancer incidence and mortality based on the amalgamated data from the aforementioned Nordic national cancer registries [3]. Alternatively, in the province of Ontario, Canada, the ICES Data Repository (formerly the Institute for Clinical Evaluative Sciences) "consists of record-level, coded and linkable health data sets that encompasses much of the publicly funded administrative health services records for the province's eligible population"[4]. This large systems-based dataset is capable of being organized and systematized on an ad hoc basis to form registries and is capable of integrating research-specific data [5–7]. Lastly, product registries are used for post-marketing surveillance in procedures, device, and pharmaceutical trials to demonstrate effectiveness and safety of products in real-world settings. There are, of course, registries that overlap categories in this simplified classification. For example, large country-wide registries for total joint arthroplasty exist in the literature that represent a procedure specific registry, but also prospectively collect robust data on the real-world effectiveness of orthopedic arthroplasty implants [8–10].

Registry level of evidence

Situating the studies that are derived from registries within the hierarchy of evidence is of importance when discussing registries. While the question of evaluation of the quality of registries will be addressed later in this paper, registries and their distinction based on assigned hierarchies of evidence are frequently used in developing guidelines and clinical decision making [1]. Many modern registries represent high-quality, prospectively designed cohort studies aimed at investigating or addressing a certain problem, hypothesis, or clinical entity, thus situating them in the realm of Level II and III quality. While very rigorously carried-out registries have been stratified just below randomized controlled trials (RCTs), the recent advent of registry-based RCTs blurs that boundary between registry-based level I and II evidence [2,11]. Registry-based RCTs represent a potential to introduce major efficiencies when it comes to patient recruitment and follow-up, while maintaining the advantages of the RCT as the best study design for investigating causal links and comparing interventions [2]. The utilization of registry-based RCTs was lauded in the New England Journal of Medicine in 2013 as "the next disruptive technology in clinical research" [12]. While the so-called registry-based RCTs are particularly advantageous as they enable rapid consecutive patient enrollment, can result in reduced per-patient cost of trial implementation, and represent pragmatic trials that can more easily be generalized to the population, the number of published studies has been slow to catch on. This may be due in part to the prerequisite requirement of a high-quality registry, which is lacking across much of the world and in many medical and surgical specialties. Ultimately, looking past the level of evidence assigned to registry-based studies, consumers of these studies must continue to understand the value of registries in delivering complementary clinical evidence that can extend the conclusions drawn from clinical trials to real-world applications of health interventions, with long-term follow-up.

Strengths of registries

The first, and most obvious, strength of registries, as alluded to above, is the sheer volume of collected data that are efficiently produced, curated, systematized, and then harnessed into scientific and clinically based conclusions. This volume and efficiency (both in time and cost) is hard to replicate with other methods. The cost-efficiency of registry-based RCTs allows the impressive pragmatic consecutive recruitment of patients [2]. One of the first such trials, dubbed the TASTE trial (Thrombus Aspiration during ST-Elevation myocardial infarction), purported to spend US $50 per patient, which was estimated at 2% the cost of a conventional RCT [13,14]. Further, there are a large number of heart-failure registry-based RCTs that purport to deliver generalizable findings at low cost, along with evidence-based and novel use of generic drugs with low costs to society [15]. Non-RCT based registries are expensive to maintain; for example the Swedish Total Joint Arthroplasty Registry costs about 35 million euros per year to maintain [16]. These costs are offset by the systems level cost-savings drawn from the data; a small reduction in the revision rates of these surgeries secondary to insights drawn from registry data can easily offset the direct registry costs to society. The prospective nature with which the data is collected also presents a unique opportunity for researchers and consumers of literature alike to draw conclusions not possible with smaller cohort or database studies.

Large registries allow us to investigate a diversity of participants that is not afforded with conventional data collection techniques and research study designs. This large catchment of population-based data helps to reduce concerns around study participation, and the various biases that evolve from research study design (i.e. volunteer bias, selection bias, etc.). The large breadth of patients included, from many different clinical contexts, adds to the external validity of the registry-based conclusions that can have a strong role in health systems planning for example; the aforementioned NORDCAN registry has been used to determine phase- and gender-specific, lifetime, and future costs of 13 cancers with high incidence levels [17]. Because of the linkage of registries to interconnected health data, especially in the context of robust EMRs, it is now possible to retrieve extensive clinical information of registry participants, thus generating thousands of data points for a single patient. Taken together, this large amount of clinical data, that is rapidly expanding, and is becoming more accessible, can generate impressive conclusions, should that 'big data' be interpreted correctly.

Considerations of registries

The large amount of data, across many registries, crossing geographic and political borders, can generate some very impressive findings and conclusions, but possess problems and hurdles as well.

Limitations of registries

The key limitation to any registry is the quality of data contained within it; not only the accuracy of the baseline characteristics, but also the consistency and rigor with which outcome measurements are generated [2,14]. The design of datasets, the methodology of data collection, and the accuracy of that data gathered by and for registries may vary across the different types and locations of registries. The quality of the data depends on the initial purpose for the registry, and the resources and methods available to the registry administrators to upkeep and clean the data. Registries can be plagued by missing or incomplete data. As well, low, or biased, or unadjudicated enrollment into the registries can bias conclusions drawn from its data. Because outcomes are usually not adjudicated in registry-based data, and follow-up is passive rather than active, outcome events may also be subject to uncertainty. For instance, in formalized trials like RCTs, patients are actively followed at pre-determined intervals, while patient follow-up data is added to registries on a less structured and pragmatic basis, or passively. As such, adverse event rates are typically lower in registries than RCTs [18,19]. Registries that choose hard clinical end points (e.g., death, or revision surgery) are thereby less susceptible to ascertainment bias and underreporting of complications or adverse events due to diverse definitions [1]. Nevertheless, at the

compromise of cost and efficiency, it is possible to adjudicate outcome events or audit data in registries to ensure certain standards data.

Registry quality assessment

Gliklich et al. [1] discuss at length the difficulty of assessing the quality of registries; in this case, quality refers both to the data and the conclusions drawn from analyses of these large swaths of registry data. The main principles in evaluating registries are founded in the principles of assessing the quality of RCTs. Ultimately, the level of quality of registry-based studies expresses the perceived confidence that the design, conduct, and analysis of the registry can be shown to protect against bias (systematic error) and errors in conclusions [20].

With that said, there are two major difficulties with assessing quality in registries. Firstly, it can often be difficult to differentiate between the quality of the design, the study conduct, and the resultant information available. Secondly, there is a lack of empirical evidence to guide the evaluation of the parameters that are alleged to indicate quality and their impact on the application and resultant utility of the evidence produced from registries [1]. A very generalized approach for assessing registry quality is to first consider whether the research question is appropriate for registry data, and whether translating those clinical questions into measurable exposures and outcomes is efficient and practical though registries. Next the data sources for registries must be evaluated as appropriate for the type of research study published (i.e. case control, cohort-based etc.). Next, the patient population included in the registry data, and the methodology of collecting data, must be representative of the clinical entity being studied. This includes the registry size as well as duration of data collection. Importantly, an appropriate comparison group for the study population must be selected from the registry data or otherwise. Lastly internal and external validity must be evaluated; the generalizability, information bias, sampling and selection bias, channeling bias (confounding by indication), loss to follow up, and an assessment of the total magnitude of bias must all be assessed. Currently, no well-defined or widely accepted quality assessment tool exists for use with registry studies.

Future directions

Global registries

To continue to broaden the data pools available to registries, and the researchers that use them, there has been a strong movement to create global registries, either by amalgamation, or by creation of novel registries. Again, looking to the domain of orthopaedics can provide us with useful case examples. The International Consortium of Orthopaedic Registries (ICOR) was created, in part, to begin the amalgamation process between various national total joint arthroplasty registries. The mandate of the ICOR included the pooling of data from existing national registries to thereby improve conclusions drawn across a generalizable global population, but also to uniform the data collection in order to facilitate more efficient exchanges [21]. For example, the consortium even created a universal bar code for more efficient data and product entry [21].

Again looking to orthopaedics, the International Orthopaedic Multicentre Study in Fracture Care (INORMUS) provides an excellent example of a de novo global registry [22–24]. With the primary objective of determining the mortality, re-operation and infection rates of musculoskeletal trauma patients within 30 days post-hospital admission, the INORMUS study continues on its process of enrolling 40,000 patients from low-to-middle income countries in Africa, Asia, and Latin America [25,26]. This move to develop a fracture registry, primary located in low-to-middle-income countries, from which the investigators will be able to draw important and clinically relevant conclusions, is an important move forward in the practice of registries. Extrapolating registries to include middle-to-low-income countries will help drive reductions in morbidity and mortality globally.

Big data

Big data is a field in computing and analytics that has grown to systematically extract or mine data and information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software. In the context of registries, the focus for the future will be on generation of these big data sets, that require dedicated computing techniques. The next generation registry will shift the database-centered thinking to a focus on integration and incorporation of layers of data, to ultimately move from surveillance to improve clinical care in real time and integration in a “big data” health information system [27].

Using the abundant national total joint registries as a case example, there is a void in the literature with respect to methodologically sound health economic analyses. The total joint replacement literature lacks up-to-date studies involving truly modern implants with appropriate follow-up, including total societal costs. National registries, and potentially international merged registries, with higher coverage and completeness have the potential to fill this void. This data can be used as a health economic instrument especially if the registries include patient-reported outcomes and can be linked to reimbursement in health care and other insurance- and societal costs [28].

One study of EMRs found that a single patient's health record was associated with a staggering average of roughly 32,000 data elements [29]. When these joint registries become linked to the EMRs of the patients included, the robust data points that are developed can further inform our literature base, and health systems. Thus, as registries become more integrated, with each other, and the medical records of its included population, they can allow the perpetual addition of patients and their data to the fold, and thus allow for more continuous and evolving conclusions to be drawn, even outside of the incremental publication of registry-based studies. It will be important in the future to ensure that the data layers that are incorporated into registries are of high-quality, and the consumers of the conclusions drawn can appropriately assess the validity of the big data.

Machine learning and big data

One subset of artificial intelligence is termed machine learning (ML); simplistically ML are computer algorithms that once built and taught the parameters of a given circumstance or problem are able to learn, reason, and self-correct without explicit programming [30]. ML algorithms are built using data from large patient databases, with the intention of finding patterns and making predictions [31]. With the foundation of registry data, inclusive of the robust EMR data now being incorporated and natural language processing based data elements, convoluted ML methods provide a novel way to derive clinical simulations, models and inform decision-making across medicine [32,33]. The future research applications of machine learning and artificial intelligence are covered elsewhere in this Research Special Edition but understanding the importance of registry-based data as the foundation of those real-time decision-making algorithms is nonetheless important.

Conclusion

Registries represent an important source of systematized data that can generate important conclusions for ultimately improving clinical care of patients. The strength of registries lies in the immense amount of dedicated data that can be harnessed to provide unique insights to large generalizable populations. Limitations for registries exist too. The importance of understanding the variability of the quality between registries cannot be understated. The future of registries is quite exciting with increases in data collection, especially with continued adoption or EMR, and application of machine learning and artificial intelligence-based algorithms to learn from the data in real time.

Conflict of Interest Statements

Luc Rubinger: This author, their immediate family, and any research foundation with which they are affiliated did not receive any financial payments or other benefits from any commercial entity related to the subject of this article. Aaron Gazendam: This author, their immediate family, and any research foundation with which they are affiliated did not receive any financial payments or other benefits from any commercial entity related to the subject of this article. Seper Ekhtiari: This author, their immediate family, and any research foundation with which they are affiliated did not receive any financial payments or other benefits from any commercial entity related to the subject of this article. Mohit Bhandari: This author, their immediate family, and any research foundation with which they are affiliated did not receive any financial payments or other benefits from any commercial entity related to the subject of this article. None of the authors have any conflict of interests to declare.

References

[1] Gliklich R.E., Leavy M.B., Dreyer N.A. (sr eds). Registries for evaluating patient outcomes: a user's guide fourth edition registries for evaluating patient outcomes: a user's guide. vol. 19(20)--H. 2020.
[2] Li G, Sajobi TT, Menon BK, Korngut L, Lowerison M, James M, et al. Registry-based randomized controlled trials- what are the advantages, challenges, and areas for future research? J Clin Epidemiol 2016;80:16–24. doi:10.1016/j.jclinepi.2016.08.003.
[3] Engholm G, Ferlay J, Christensen N, Bray F, Gjerstorff ML, Klint A, et al. NORDCAN–a Nordic tool for cancer information, planning, quality control and research. Acta Oncol 2010;49:725–36. doi:10.3109/02841861003782017.
[4] ICES. ICES Data n.d.
[5] Axelrod D, Veljkovic A, Zochowski T, Marks P, Mahomed N, Wasserstein D. Risk of ankle fusion or arthroplasty after operatively and nonoperatively treated ankle fractures: a matched cohort population study. J Orthop Trauma 2020;34.
[6] Pincus D, Ravi B, Wasserstein D, Huang A, Paterson JM, Nathens AB, et al. Association between wait time and 30-day mortality in adults undergoing hip fracture surgery. JAMA 2017;318:1994–2003. doi:10.1001/jama.2017.17606.
[7] Bozzo A, Ekhtiari S, Madden K, Bhandari M, Ghert M, Khanna V, et al. Incidence and predictors of prosthetic joint infection following primary total knee arthroplasty: a 15-year population-based cohort study. J Arthroplasty 2021. doi:10.1016/j.arth.2021.10.006.
[8] Graves SE, Davidson D, Ingerson L, Ryan P, Griffith EC, McDermott BFJ, et al. The Australian orthopaedic association national joint replacement registry. Med J Aust 2004;180:S31–4. doi:10.5694/j.1326-5377.2004.tb05911.x.
[9] Bohm ER, Dunbar MJ, Bourne R. The Canadian joint replacement registry-what have we learned? Acta Orthop 2010;81:119–21. doi:10.3109/17453671003685467.
[10] Springer BD, Levine BR, Golladay GJ. Highlights of the 2020 american joint replacement registry annual report. Arthroplast Today 2021;9:141–2. doi:10.1016/j.artd.2021.06.004.
[11] Leta TH, Gjertsen J-E, Dale H, Hallan G, Lygre SHL, Fenstad AM, et al. Antibiotic-loaded bone cement in prevention of periprosthetic joint infections in primary total knee arthroplasty: a register-based multicentre randomised controlled non-inferiority trial (ALBA trial). BMJ Open 2021;11:e041096. doi:10.1136/bmjopen-2020-041096.
[12] Lauer MS, D'Agostino RB. The randomized registry trial — the next disruptive technology in clinical research? N Engl J Med 2013;369:1579–81. doi:10.1056/NEJMp1310102.
[13] Fröbert O, Lagerqvist B, Gudnason T, Thuesen L, Svensson R, Olivecrona GK, et al. Thrombus Aspiration in ST-Elevation myocardial infarction in Scandinavia (TASTE trial). A multicenter, prospective, randomized, controlled clinical registry trial based on the Swedish angiography and angioplasty registry (SCAAR) platform. Study design and. Am Heart J 2010;160:1042–8. doi:10.1016/j.ahj.2010.08.040.
[14] James S, Rao SV, Granger CB. Registry-based randomized clinical trials—A new clinical trial paradigm. Nat Rev Cardiol 2015;12:312–16. doi:10.1038/nrcardio.2015.33.
[15] Lund LH, Oldgren J, James S. Registry-Based Pragmatic Trials in Heart Failure: current Experience and Future Directions. Curr Heart Fail Rep 2017;14:59–70. doi:10.1007/s11897-017-0325-0.
[16] Delaunay C. Registries in orthopaedics. Orthop Traumatol Surg Res 2015;101:S69–75. doi:10.1016/j.otsr.2014.06.029.
[17] Bugge C, Brustugun OT, Sæther EM, Kristiansen IS. Phase- and gender-specific, lifetime, and future costs of cancer: a retrospective population-based registry study. Medicine (Baltimore) 2021;100:e26523. doi:10.1097/MD.0000000000026523.
[18] Phillips R, Sauzet O, Cornelius V. Statistical methods for the analysis of adverse event data in randomised controlled trials: a scoping review and taxonomy. BMC Med Res Methodol 2020;20:288. doi:10.1186/s12874-020-01167-9.
[19] Pitrou I, Boutron I, Ahmad N, Ravaud P. Reporting of safety results in published reports of randomized controlled trials. Arch Intern Med 2009;169:1756–61. doi:10.1001/archinternmed.2009.306.
[20] Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Control Clin Trials 1995;16:62–73. doi:10.1016/0197-2456(94)00031-w.
[21] Sedrakyan A, Paxton EW, Marinac-Dabic D. Stages and Tools for Multinational Collaboration: the Perspective from the Coordinating Center of the International Consortium of Orthopaedic Registries (ICOR). JBJS 2011:93.
[22] Sprague S, McKay P, Li CS, Ivers R, Moroz PJ, Jagnoor J, et al. International Orthopaedic Multicenter Study in Fracture Care: coordinating a Large-Scale Multicenter Global Prospective Cohort Study. J Orthop Trauma 2018;32(Suppl 7):S58–63. doi:10.1097/BOT.0000000000001287.
[23] Pouramin P, Li CS, Sprague S, Busse JW, Bhandari M. A multicenter observational study on the distribution of orthopaedic fracture types across 17 low- and middle-income countries. OTA Int Open Access J Orthop Trauma 2019;2:e026. doi:10.1097/OI9.0000000000000026.
[24] INternational ORthopaedic MUlticentre Study (INORMUS) in Fracture Care: protocol for a Large Prospective Observational Study. J Orthop Trauma 2015;29(Suppl 1):S2–6. doi:10.1097/BOT.0000000000000404.
[25] Foote CJ, Mundi R, Sancheti P, Gopalan H, Kotwal P, Shetty V, et al. Musculoskeletal trauma and all-cause mortality in India: a multicentre prospective cohort study. Lancet 2015;385 Suppl:S30. doi:10.1016/S0140-6736(15)60825-X.
[26] Pouramin P, Li CS, Busse JW, Sprague S, Devereaux PJ, Jagnoor J, et al. Delays in hospital admissions in patients with fractures across 18 low-income and middle-income countries (INORMUS): a prospective observational study. Lancet Glob Heal 2020;8:e711–20. doi:10.1016/S2214-109X(20)30067-X.
[27] Mate S, Köpcke F, Toddenroth D, Martin M, Prokosch H-U, Bürkle T, et al. Ontology-based data integration between clinical and research systems. PLoS ONE 2015;10:e0116656. doi:10.1371/journal.pone.0116656.
[28] Malchau H, Garellick G, Berry D, Harris WH, Robertson O, Kärrlholm J, et al. Arthroplasty implant registries over the past five decades: development, current, and future impact. J Orthop Res 2018;36:2319–30. doi:10.1002/jor.24014.
[29] Milinovich A, Kattan MW. Extracting and utilizing electronic health data from Epic for research. Ann Transl Med 2018;6:42. doi:10.21037/atm.2018.01.13.
[30] Ben-Israel D, Jacobs WB, Casha S, Lang S, Ryu WHA, de Lotbiniere-Bassett M, et al. The impact of machine learning on patient care: a systematic review. Artif Intell Med 2020;103:101785. doi:10.1016/j.artmed.2019.101785.
[31] Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA 2018;319:1317–18. doi:10.1001/jama.2017.18391.
[32] Waring J, Lindvall C, Umeton R. Automated machine learning: review of the state-of-the-art and opportunities for healthcare. Artif Intell Med 2020;104:101822. doi:10.1016/j.artmed.2020.101822.
[33] Liu Y, Chen P-HC, Krause J, Peng L. How to Read Articles That Use Machine Learning: users' Guides to the Medical Literature. JAMA 2019;322:1806–16. doi:10.1001/jama.2019.16489.