REVIEW

Opportunities for the use of routinely collected data for the generation of large randomized evidence in Colombia [version 1; peer review: 2 approved with reservations]

Jonny Alejandro Garcia Luna ^{1,2}, Eduardo López-Medina^{1,3,4},
Norman Danilo Maldonado-Vargas ⁵, Adrian D. Smith ⁶

¹Centro Internacional de Entrenamiento e Investigaciones Médicas (CIDEIM), Cali, Colombia

²Universidad Icesi, Cali, Colombia

³Centro de Estudios en Infectología Pediátrica (CEIP), Cali, Colombia

⁴Department of pediatrics, Universidad del Valle, Cali, Colombia

⁵Centro de Estudios en Protección Social y Economía de la Salud (PROESA), Universidad Icesi, Cali, Colombia

⁶Nuffield Department of Population Health, University of Oxford, Oxford, UK

v1 First published: 15 Jul 2021, 6:182
<https://doi.org/10.12688/wellcomeopenres.17036.1>
Latest published: 15 Jul 2021, 6:182
<https://doi.org/10.12688/wellcomeopenres.17036.1>

Abstract

Randomized clinical trials are the cornerstone design for the evaluation of the safety and efficacy of health interventions. Furthermore, morbidity and mortality rates could be reduced if evidence of better interventions is sought and used to inform medical practice. However, only small to moderate, yet worthwhile, effects can be expected from such interventions. Therefore, moderate random error and moderate biases must be avoided during the design, conduct and analysis of trials. Routinely collected data, such as vital statistics, hospital episode statistics and surveillance data, could be used to enhance recruitment and follow-up a large number of patients, reducing both random error and moderate biases. Here, we discuss the opportunities and challenges for the use of these data for clinical studies in Colombia.

Keywords

Clinical trials, routinely collected health data, disease registries

Open Peer Review

Approval Status

	1	2
version 1 15 Jul 2021	view	view
1. Diego Fernando Rojas-Gualdron , CES University, Medellín, Colombia		
2. Alasdair Henderson , London School of Hygiene and Tropical Medicine, London, UK		
Any reports and responses or comments on the article can be found at the end of the article.		

Corresponding author: Jonny Alejandro Garcia Luna (jagarcia@cideim.org.co)

Author roles: **Garcia Luna JA:** Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; **López-Medina E:** Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Maldonado-Vargas ND:** Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Smith AD:** Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome [214641, <https://doi.org/10.35802/214641>]

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2021 Garcia Luna JA *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Garcia Luna JA, López-Medina E, Maldonado-Vargas ND and Smith AD. **Opportunities for the use of routinely collected data for the generation of large randomized evidence in Colombia [version 1; peer review: 2 approved with reservations]** Wellcome Open Research 2021, 6:182 <https://doi.org/10.12688/wellcomeopenres.17036.1>

First published: 15 Jul 2021, 6:182 <https://doi.org/10.12688/wellcomeopenres.17036.1>

Introduction

Over the last two centuries life expectancy has improved remarkably, rising from less than 50 years in the 19th century to more than 70 years in 2010^{1,2}. Improved sanitation, hygiene, nutrition, the development of vaccines and the advancement of medical knowledge have been the main drivers of the increased longevity³. However, millions of premature deaths could be avoided if evidence of better interventions is sought and used to inform medical practice⁴. The leading causes of morbidity and mortality worldwide are cardiovascular diseases (CVD) and cancer⁵, both of which are multifactorial diseases with complex pathophysiological processes leading to disability and mortality^{6,7}. Thus, interventions targeting only one or a reduced number of these processes can only be expected to produce small to moderate effects on clinical outcomes. Nevertheless, every year around 18 million deaths are attributed to CVD⁵; even a small reduction of 10% in CVD mortality could potentially avoid hundreds of thousands of deaths.

On the other hand, although large effects of antimicrobial treatments are expected in infectious diseases, once an effective treatment is available, the optimisation of therapeutic regimens or the evaluation of interventions for resistant microorganisms may require also large studies. For example, only 107 patients were necessary to demonstrate the efficacy of streptomycin for the treatment of tuberculosis in 1950⁸, while thousands of patients have been randomized to evaluate the efficacy of shorter regimens for this disease⁹. Furthermore, the randomization of thousands of patients with coronavirus disease 2019 (COVID-19) in the RECOVERY and SOLIDARITY trials was necessary to offer robust evidence of the presence (dexamethasone¹⁰ and tocilizumab^(RECOVERY)) or absence (hydroxychloroquine^{11,12} and lopinavir/ritonavir^{11,13}) of therapeutic efficacy.

If only small to moderate effects can be expected from health interventions, moderate random error and biases must be avoided during the design, conduct and analysis of clinical trials^{4,14}. Otherwise, the magnitude of such errors on the estimations of efficacy may obscure or completely reverse the estimated effect of the interventions under evaluation^{4,14}. The randomization of a large number of participants in controlled clinical trials reduces both random error and biases¹⁵. Furthermore, both large observational studies and small randomized clinical trials may produce chance or biased results, hindering the implementation of health interventions in clinical practice⁴. For example, while a nationwide study using routinely collected healthcare data (RCHD) of patients with cancer in Denmark found a non-causal association between statin use and an increased risk of vascular mortality (hazard ratio [HR] 1.25, 95% confidence interval [95%CI] 1.21-1.41)¹⁶, several meta-analyses of randomized clinical trials have shown a 21% lower risk of vascular events for every 1 mmol/L reduction of low-density lipoprotein cholesterol (RR 0.79, 95%CI 0.77 – 0.83)¹⁷. Furthermore, moderate random error and biases may produce chance findings or increase the effect size of futile interventions; for example, a meta-analysis of

small clinical trials of magnesium sulphate versus placebo suggested a reduction of about a half in the odds of death (odds ratio [OR] 0.45, $p < 0.001$), without evidence of heterogeneity in study results¹⁸. In contrast, the ISIS-4 clinical trial that randomized about 58,000 patients to receive either sulphate magnesium or placebo, proved that it was of little value for the prevention of vascular mortality after an acute myocardial infarction (OR 1.06, $p = 0.07$)¹⁹.

Colombia, with a population of almost 50 million people, near universal health care coverage and a diverse range of diseases²⁰, is a perfect niche for the implementation of large randomized clinical trials including thousands of participants. Furthermore, the development of technological innovations and powerful computational systems has allowed for the routine electronic collection of health care data for administrative and surveillance purposes^{20,21}. Here, we discuss how RCHD can be used in the design and conduct of large randomized clinical trials and what the major challenges and opportunities are for these in Colombia. [Table 1](#) summarizes selected clinical trials that included RCHD in their designs.

Discussion

Enhancing trial recruitment using routinely collected data

About two thirds of clinical trials do not achieve their target sample size and often require extensions and budget supplements²². The use of RCHD in the planning and conduct of clinical trials may increase the recruitment of participants and reduce the time required to achieve the sample size, since using RCHD is an efficient method for selecting study sites and identifying potential trial participants meeting eligibility criteria. For example, surveillance data could be used to identify the transmission patterns of vector-borne diseases and inform the selection of the most suitable communities for vaccine trials or other control strategies²³. Similarly, data from cancer registries could be used to identify the sites with the highest potential for recruitment based on the number of patients who met inclusion criteria and who attend the potential clinical sites per year, reducing the overall cost of the study and increasing its efficiency, since less time and fewer sites would be required^{24,25}. Furthermore, monitoring the recruitment process in clinical trials has been identified as a key component for the optimal conduct of trials²⁶, and RCHD such as hospital episode statistics, disease registries and surveillance data could be used to define recruitment goals and compare such goals against potential candidates. For example, if a site reports the screening of 100 participants, but surveillance data indicates that more than 1,000 cases were seen in the site, the reasons for the poor recruitment could be explored and addressed. However, this approach requires timely data coding, cleaning and distribution^{21,24}. Finally, RCHD provides a contemporary sampling frame that could be used to identify potentially eligible trial participants. The ASCEND trial, one of the largest primary prevention studies in diabetic patients, enrolled 15,480 participants, 58% of whom (9,013) were identified from central diabetes registers²⁷.

Table 1. Clinical trials using routinely collected healthcare data (RCHD) in their design.

Trial	Country	Sample size	Description	Use of RCHD
SAFE-PCI ^{24,28}	USA	1,787	Randomized open label clinical trial to compare the safety of radial and femoral access for percutaneous coronary intervention	National Cardiovascular Data Registry was used to optimize the selection of sites according to available data on patient volume and acceptable experience of each site.
ASCEND ^{27,29–31}	UK	15,480	2x2 factorial design randomized clinical trial to assess the efficacy to prevent cardiovascular events or cancer in people with diabetes, of aspirin 100mg daily versus placebo and omega 3 fatty acids 1g daily versus placebo.	Identification of participants from central diabetes registers. Follow-up supplemented by data linkage to central databases (hospital episode statistics, cancer registers and death registers).
DETO2X ^{32,33}	Sweden	6,629	A randomized controlled clinical trial of oxygen supplementation in patients with suspected AMI and normal oxygen levels	Identification of participant from a central diseases registry. Follow-up through diseases register and national vital statistics.
TASTE ³⁴	Sweden	7,244	A randomized controlled clinical trial of thrombus aspiration during percutaneous coronary intervention	Identification of participant from a central diseases registry. Follow-up through diseases register and national vital statistics.
Swedish clinical trials of mammography ³⁵	Sweden	>200,000	Randomized controlled clinical trial of mammography.	Extended follow up through vital statistics.
AgeX (AgeX)	UK	>4 million	Randomized controlled clinical trial of extended age range of mammography (47 to 73 years).	Follow-up through central databases.

UK: United Kingdom, USA: United States of America

Quantifying representativeness using routinely collected data

The generalizability of trial results is one of the most important considerations for clinicians, researchers and policy makers⁴. However, people who agree to participate in a clinical trial are different from people who are excluded or do not agree to participate³⁶. RCHD may aid in the quantification of important differences between trial participants and the rest of potential participants with the diseases. For example, data from diseases registries or surveillance databases could be used to compare clinical characteristics of trial participants and other people with the disease³⁶, highlighting potential knowledge gaps and differences in potential effect modifiers that compromise the generalizability of the results.

Passive follow-up of trial participants using routinely collected data

Bias introduced by selective loss to follow-up is a major threat to the potential validity of clinical trials¹⁴, particularly when it is associated with either the intervention or the outcomes. Hard outcomes such as mortality, hospitalization and reportable communicable diseases are routinely collected and may be used to complement efficacy and safety outcomes³⁰. However, mild events such as mild bleeding or non-severe muscle pain are less likely to trigger a health care visit, and therefore soft outcomes are unlikely to be obtained from RCHD. Furthermore, the evaluation of new drugs and devices usually requires a

comprehensive safety report and serial blood sampling for analysis of pharmacokinetics and pharmacodynamics, hence active follow-up remains as an essential component of trial conduction³⁶. Nevertheless, RCHD has successfully supplemented outcome information in several clinical trials, reducing loss to follow-up for the most important outcomes^{29,31,33,34}. In the ASCEND trial, outcome data was supplemented by data linkage to central databases (hospital episode statistics, cancer registries and vital statistics) obtaining 99.1% of complete follow up³⁰.

On the other hand, RCHD could be used to extend the follow-up of trial participants beyond the scheduled trial end and estimate long term efficacy and unintended consequences of health interventions³⁵. In the Swedish clinical trials of mammography, the average trial duration was 6 years, after which control participants were invited for screenings³⁵. However, the use of vital statistics supplemented the follow-up of more than 200,000 women over an average of more than 15 years after the scheduled trial end, allowing the inclusion of more than 1,000 incident breast cancer deaths and finding a significant 21% reduction in breast cancer mortality associated with the invitation to mammographic screening (RR=0.79, 95%CI 0.70-0.89)³⁵.

Furthermore, RCHD could be used to monitor very rare side effects of health interventions once they have been rolled out,

for example the potential association between cerebral vein thrombosis and the AstraZeneca COVID-19 vaccine³⁷.

The cost-effectiveness of using routinely collected data for clinical trials

The regulatory framework surrounding the conduct of clinical trials has increased their cost to almost prohibitive levels³⁸. However, the use of RCHD to optimize trial site selection, recruitment and follow-up, not only aids in achieving large sample sizes and limiting loss to follow-up and therefore to reduce moderate random error and bias¹⁵, it also helps to reduce trial costs significantly^{27,36}. While a typical large randomized clinical trial costs hundreds of millions of US dollars³⁸, the overall cost of the ASCEND trial was less than 10 million US dollars²⁷ and the use of the SWEDHEART registry reduced the cost of the TASTE trial by more than 90%³⁶. Additionally, the implementation of streamline procedures and the use of RCHD have allowed the inclusion of more than 4 million women in the [AgeX](#) trial, the largest randomized control clinical trial ever conducted.

Furthermore, the linkage of clinical trial data with RCHD allows the implementation of the adaptive trials with multiples arms, including at least one control arm as well as in large cluster randomized clinical trials. For example, linkage through healthcare databases is expected to provide several years of follow-up of participants in the [RECOVERY](#) trial.

Potential problems with the use of routinely collected data

The acceptability of using RCHD to either identify potential trial participants or for passive follow-up is an important issue³⁹. In the ASCEND trial, the use of identifiable data raised concerns in a small number of participants who were unaware of the legislation that authorized the use of their data, however most participants were satisfied after a short explanation²⁷. Similarly, a survey of 589 oncological patients in Ontario, Canada found that most patients (93%) were willing to authorize long-term linkage of their personal information, although they would prefer active follow-up through clinic appointments⁴⁰. However, long-term data linkage generates the problem of participants who may not be aware that they are still being followed-up and thus are unlikely to request their withdrawal from the trial, particularly in surgical trials where the intervention is performed in only one visit.

A second potential problem arising from the use of RCHD for the follow-up of trial participants is a potential lower compliance with the allocated treatment, due to the decreased personal interaction with research staff and reduced participant's engagement, although this potentially allows the reliable estimation of the real-world effectiveness of health interventions. In the ASCEND trial, where participants were followed up using mailed questionnaires and RCHD, there was an adherence of 70% over a mean follow-up time of 7.4 years²⁹. Similarly, in the ARRIVE trial, also a primary prevention study, the compliance was 80% over a mean follow-up time of 5 years, despite the active, face-to-face follow-up of trial

participants⁴¹. Thus, the use of RCHD has had little impact on long term adherence to the allocated treatment, and data linkage is a suitable method to supplement outcome measures in large randomized clinical trials, particularly for surgical or invasive interventions, where practically all participants received the allocated interventions^{28,34}.

A third potential problem with the use of RCHD is the quality of the data since the original purpose of RCHD was not to supplement outcome ascertainment in clinical trials. However, several validation studies have shown that hard outcomes, such as cause-specific mortality, can be reliably collected using RCHD^{42,43}. In a validation study conducted in Sweden, the cause of death reported in the central vital statistics was consistent with an independent endpoint committee in about 93% of the cases⁴². A similar study, conducted in Scotland, found that linkage to electronic health records was reliable since out of 58 recorded deaths there was an exact match on the dates and cause of death in 56 cases, as well as generally good agreement over a thousand hospital admissions⁴³. However, the use of hard outcomes, which are rare and dichotomous, paradoxically requires an increase in the required sample size to yield enough power to detect the effect of the intervention. On the other hand, in a sample of the Million Women Study in the United Kingdom, the diagnosis of dementia was first mentioned in primary health care medical records, an average of 1.6 years earlier than in hospital episode statistics⁴⁴, suggesting that primary care databases could be an important source of information for the timely ascertainment of outcomes. Nevertheless, misclassification of outcomes collected from RCHD are likely to be non-differential and therefore may have a low impact on the estimations of efficacy as long as the sample size is large enough⁴⁵.

A major disadvantage of using RCHD is the inconsistent quality of linkage methods. Linked data is a mix of correctly matched, mistakenly linked and incorrectly unmatched records⁴⁶. The two most commonly used methods for data linkage are deterministic and probabilistic⁴⁶. Deterministic methods only match records when they agree exactly on the linkage variable, thus this method is preferred when a unique identifier, such as an ID number or social insurance number, are available across databases⁴⁶. This method yields the highest specificity⁴⁶, however a unique identifier is not available in all settings and other linkage methods should be implemented. Probabilistic methods match records based on several variables that may not be unique, are prone to spelling mistakes and that may change over time such as name, date of birth and address. During probabilistic matching, the probability of belonging to a certain participant is assigned to each record and only records with a probability higher than a pre-specified threshold are matched⁴⁶. Unfortunately, the linkage method is outside the control of researchers since the algorithms used are decided and implemented by the institution that held the databases (NHS). However, misclassification of outcome measures is expected to be non-differential and its impact on the estimations of efficacy is expected to be low as long as the trial is large enough⁴⁵.

Challenges and opportunities for the use of routinely collected data for trials in Colombia

Colombia's Health System (CHS) works mainly under a social health insurance scheme that finances and manages healthcare through risk pooling, while a small segment of the population with the highest income acquire private health insurance to reduce the time to access to health services. The two principal insurance schemes are (i) the contributory regime for the formal workers and their families who make monthly financial contributions to the system through employer and employee payroll contributions, and (ii) the subsidized regime for informal workers who do not contribute to the system. Both schemes fairly cover the same services, including but not limited to outpatient, emergency and hospitalization care and most medications⁴⁷. The system is regulated and overseen by the Ministry of Health (MoH), and healthcare services are delivered by public and private health institutions known as IPS (Instituciones Prestadoras de Servicios [Health Service Provider Institutions]), financed through a health insurance premium called UPC (Unidad de Pago por Capitación [Per capita payment]), received from the MoH by health insurance companies known as EPS (Empresas Promotoras de Salud [Health promotion Companies]), and funded with a combination of public budget and payroll taxes. In order to regulate the use of healthcare, users from the contributory regime and some of the subsidized regime are in charge of paying moderating fees and/or small co-payments. RCHD is collected mainly by the IPS and EPS, and most of that information must be reported to the Ministry of Health in the Integrated System of Information on Social Protection (SISPRO), a centralized system of information for data management. In addition, the National Department of Statistics (DANE) administers a set of nationally representative surveys that provide information on health status and on determinants of health, as well as vital statistics.

At least two prerequisites for the use of RCHD in the design and conduction of clinical trials can be inferred from the experience in high income countries. First, the information must be collected and available, and secondly, clear legislation and protocols for its use must be in place. Regarding data collection, Colombia's MoH has SISPRO, which centralizes the administrative data on health, healthcare and social protection as well as national surveys related to health (e.g. demographic and health surveys)⁴⁸. The system includes, among others, nationwide registries of morbidity and mortality such as the system of public health surveillance (SIVIGILA), consisting of data on reportable events collected by the health surveillance system (INS), and the system of civil registry and vital statistics (SRECV) consisting all births and all deaths by causes (DANE). Furthermore, the national registry of patients with high healthcare cost (CAC), includes the record of clinical characteristics of thousands of patients with HIV, C hepatitis, end-stage renal disease, cancer, rheumatoid arthritis and hemophilia, and the system of information on provision of healthcare (RIPS) aggregate data on millions of

consultations, procedures and diagnostic tests provided under the public insurance health system²⁰. All data sources can potentially be combined using the national identification number as the unique identifier.

The data takes time in flowing from primary collection to availability in SISPRO. In general, primary collection of health and healthcare events occurs in the healthcare provider, who sends reports to other actors in the health system due to either a direct legal responsibility to inform public health surveillance systems and civil registry systems, or an economic and institutional arrangement with the payer (health insurance companies). Depending on the system of information and the internal processes of validation, the flow of data can take as long as two years to become available for consultation at SISPRO. For example, while compulsory reporting cases are reported to the surveillance system by the IPS in a weekly basis, the data requires several months for cleaning and processing and is often analyzed quarterly (INS). RIPS are reported monthly by the IPS, while the CAC collects data yearly. Delays to permit validation have been shown to improve reliability to some of the data available at SISPRO such as the public health surveillance data^{49,50} and death certificates⁵¹.

Conclusions

The availability of RCHD suggests that the next step is to conduct validation analysis of SISPRO data to ascertain trial outcomes and inform trial design. Furthermore, timeliness of data availability could be improved by working together with the MoH and the main actors of the health system (providers and payers) to access the data in shorter times by making clinical trials part of healthcare risk management, one of the main structural functions of the health system. Regarding legislation and protocols, there is legislation for the use of RCHD for research purposes (law 1266/ 2008), however to our knowledge, there is no clear legislation or protocols in Colombia to use RCHD to inform trial design. Furthermore, the acceptability of using RCHD for the identification of potential trial participants and for long-term data linkage have not been determined in our population. Notwithstanding, a recent evaluation of health consequences of aerial spraying of glyphosate used RIPS from aspersión areas to estimate the effects of this intervention on the consultation pattern in the affected populations²⁰. Therefore, the use of RCHD is feasible, but a clear regulation and expeditious procedures are urgently needed for the routine use of this data for health research, especially for clinical trials.

Data availability

No data are associated with this article.

Acknowledgements

With their confirmed permission, the authors would like to thank Dr. David Rebellon and Alexandra Cossio for their review and comments on an initial version of this manuscript.

References

1. Wilmoth JR: **Demography of longevity: Past, present, and future trends.** *Exp Gerontol.* 2000; **35**(9–10): 1111–29.
[PubMed Abstract](#) | [Publisher Full Text](#)
2. Norheim OF, Jha P, Admasu K, *et al.*: **Avoiding 40% of the premature deaths in each country, 2010-30: Review of national mortality trends to help quantify the UN Sustainable Development Goal for health.** *Lancet.* 2015; **385**(9964): 239–52.
[PubMed Abstract](#) | [Publisher Full Text](#)
3. Peto R, Lopez AD, Norheim OF: **Halving premature death.** *Science.* 2014; **345**(6202): 1272.
[PubMed Abstract](#) | [Publisher Full Text](#)
4. Peto R, Collins R, Gray R: **Large scale Randomized evidence: Large simple trials and overviews of trials.** *J Clin Epidemiol.* 1995; **48**(1): 23–40.
[PubMed Abstract](#) | [Publisher Full Text](#)
5. GBD 2017 Causes of Death Collaborators: **Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017.** *Lancet.* 2018; **392**(10159): 1736–88.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Libby P, Buring JE, Badimon L, *et al.*: **Atherosclerosis.** *Nat Rev Dis Prim.* 2019; **5**(1): 56.
[PubMed Abstract](#) | [Publisher Full Text](#)
7. Harbeck N, Penault-Llorca F, Cortes J, *et al.*: **Breast cancer.** *Nat Rev Dis Primers.* 2019; **5**(1): 66.
[PubMed Abstract](#) | [Publisher Full Text](#)
8. Marshall G, Blacklock JWS, Cameron C, *et al.*: **Streptomycin in the treatment of pulmonary tuberculosis.** *Br J Tuberc Dis Chest.* 1950; **44**(4): 98–104.
[Publisher Full Text](#)
9. Imperial MZ, Nahid P, Phillips PPJ, *et al.*: **A patient-level pooled analysis of treatment-shortening regimens for drug-susceptible pulmonary tuberculosis.** *Nat Med.* 2018; **24**(11): 1708–1715.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
10. Recovery collaborative group, Horby P, Lim WS: **Dexamethasone in Hospitalized Patients with Covid-19.** *N Engl J Med.* 2021; **384**(8): 693–704.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
11. WHO Solidarity trial consortium, Pan H, Peto R: **Repurposed Antiviral Drugs for Covid-19 — Interim WHO Solidarity Trial Results.** *N Engl J Med.* 2021; **384**(6): 497–511.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
12. Recovery collaborative group, Horby P, Mafham M: **Effect of Hydroxychloroquine in Hospitalized Patients with Covid-19.** *N Engl J Med.* 2020; **383**(21): 2030–40.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
13. RECOVERY Collaborative Group: **Lopinavir-ritonavir in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial.** *Lancet.* 2020; **396**(10259): 1345–52.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Collins R, MacMahon S: **Reliable assessment of the effects of treatment on mortality and major morbidity.** *Clinical trials.* *Lancet.* 2001; **357**(9253): 373–80.
[PubMed Abstract](#) | [Publisher Full Text](#)
15. Collins R, Bowman L, Landray M, *et al.*: **The Magic of Randomization versus the Myth of Real-World Evidence.** *N Engl J Med.* 2020; **382**(7): 674–678.
[PubMed Abstract](#) | [Publisher Full Text](#)
16. Nielsen SF, Nordestgaard BG, Bojesen SE: **Statin use and reduced cancer-related mortality.** *N Engl J Med.* 2013; **367**(19): 1792–802.
[PubMed Abstract](#) | [Publisher Full Text](#)
17. Collins R, Reith C, Emberson J, *et al.*: **Interpretation of the evidence for the efficacy and safety of statin therapy.** *Lancet.* 2016; **388**(10059): 2532–2561.
[PubMed Abstract](#) | [Publisher Full Text](#)
18. Teo KK, Yusuf S, Collins R, *et al.*: **Effects of intravenous magnesium in suspected acute myocardial infarction: Overview of randomised trials.** *BMJ.* 1991; **303**(6816): 1499–503.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
19. ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group *: **ISIS-4: A randomised factorial trial assessing early oral aspirin, oral mononitrate, and intravenous magnesium sulphate in 58,050 patients with suspected acute myocardial infarction.** *Lancet.* 1995; **345**(8951): 669–85.
[PubMed Abstract](#) | [Publisher Full Text](#)
20. Camacho A, Mejía D: **The health consequences of aerial spraying illicit crops: The case of Colombia.** *J Health Econ.* 2017; **54**: 147–160.
[PubMed Abstract](#) | [Publisher Full Text](#)
21. Bohensky MA, Jolley D, Sundararajan V, *et al.*: **Data Linkage: A powerful research tool with potential problems.** *BMC Health Serv Res.* 2010; **10**: 346.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
22. McDonald AM, Knight RC, Campbell MK, *et al.*: **What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies.** *Trials.* 2006; **7**: 9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
23. Mwangoka G, Ogutu B, Msambichaka B, *et al.*: **Experience and challenges from clinical trials with malaria vaccines in Africa.** *Malar J.* 2013; **12**(1): 86.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
24. Hess CN, Rao SV, Kong DF, *et al.*: **Embedding a randomized clinical trial into an ongoing registry infrastructure: Unique opportunities for efficiency in design of the Study of Access site for Enhancement of Percutaneous Coronary Intervention for Women (SAFE-PCI for Women).** *Am Heart J.* 2013; **166**(3): 421–428.e1.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
25. Hurtado-Chong A, Joeris A, Hess D, *et al.*: **Improving site selection in clinical studies: A standardised, objective, multistep method and first experience results.** *BMJ Open.* 2017; **7**(7): e014796.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
26. Huang GD, Bull J, Johnston McKee K, *et al.*: **Clinical trials recruitment planning: A proposed framework from the Clinical Trials Transformation Initiative.** *Contemp Clin Trials.* 2018; **66**: 74–9.
[PubMed Abstract](#) | [Publisher Full Text](#)
27. Aung T, Haynes R, Barton J, *et al.*: **Cost-effective recruitment methods for a large randomised trial in people with diabetes: A Study of Cardiovascular Events IN Diabetes (ASCEND).** *Trials.* 2016; **17**(1): 286.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
28. Rao SV, Hess CN, Barham B, *et al.*: **A registry-based randomized trial comparing radial and femoral approaches in women undergoing percutaneous coronary intervention: The SAFE-PCI for women (study of access site for enhancement of PCI for women) trial.** *JACC Cardiovasc Interv.* 2014; **7**(8): 857–67.
[PubMed Abstract](#) | [Publisher Full Text](#)
29. ASCEND Study Collaborative Group, Bowman L, Mafham M, *et al.*: **Effects of aspirin for primary prevention in persons with diabetes mellitus.** *N Engl J Med.* 2018; **379**(16): 1529–39.
[PubMed Abstract](#) | [Publisher Full Text](#)
30. Bowman L, Mafham M, Stevens W, *et al.*: **ASCEND: A Study of Cardiovascular Events IN Diabetes: Characteristics of a randomized trial of aspirin and of omega-3 fatty acid supplementation in 15,480 people with diabetes.** *Am Heart J.* 2018; **198**: 135–44.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
31. The ASCEND study collaborative group: Bowman L, Mafham M, *et al.*: **Effects of n-3 Fatty Acid Supplements in Diabetes Mellitus.** *N Engl J Med.* 2018; **379**(16): 1540–50.
[PubMed Abstract](#) | [Publisher Full Text](#)
32. Hofmann R, James SK, Svensson L, *et al.*: **DETermination of the role of OXYgen in suspected Acute Myocardial Infarction trial.** *Am Heart J.* 2014; **167**(3): 322–8.
[PubMed Abstract](#) | [Publisher Full Text](#)
33. Hofmann R, James SK, Jernberg T, *et al.*: **Oxygen therapy in suspected acute myocardial infarction.** *N Engl J Med.* 2017; **377**(13): 1240–9.
[PubMed Abstract](#) | [Publisher Full Text](#)
34. Fröbert O, Lagerqvist B, Gudnason T, *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 rationale.** *Am Heart J.* 2010; **160**(6): 1042–8.
[PubMed Abstract](#) | [Publisher Full Text](#)
35. Nyström L, Andersson I, Björkstam N, *et al.*: **Long-term effects of mammography screening: Updated overview of the Swedish randomised trials.** *Lancet.* 2002; **359**(9310): 909–19.
[PubMed Abstract](#) | [Publisher Full Text](#)
36. James S, Rao SV, Granger CB: **Registry-based randomized clinical trials--A new clinical trial paradigm.** *Nat Rev Cardiol.* 2015; **12**(5): 312–6.
[PubMed Abstract](#) | [Publisher Full Text](#)
37. Pottegård A, Lund LC, Karlstad Ø, *et al.*: **Arterial events, venous thromboembolism, thrombocytopenia, and bleeding after vaccination with Oxford-AstraZeneca ChAdOx1-S in Denmark and Norway: Population based cohort study.** *BMJ.* 2021; **373**: n1114.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
38. Eisenstein EL, Collins R, Cracknell BS, *et al.*: **Sensible approaches for reducing clinical trial costs.** *Clin Trials.* 2008; **5**(1): 75–84.
[PubMed Abstract](#) | [Publisher Full Text](#)
39. Davies C, Collins R: **Balancing potential risks and benefits of using confidential data.** *BMJ.* 2006; **333**(7563): 349–51.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
40. Hay AE, Leung YW, Pater JL, *et al.*: **Linkage of clinical trial and administrative data: A survey of cancer patient preferences.** *Curr Oncol.* 2017; **24**(3): 161–7.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
41. Gaziano JM, Brotons C, Coppolecchia R, *et al.*: **Use of aspirin to reduce risk of initial vascular events in patients at moderate risk of cardiovascular disease (ARRIVE): a randomised, double-blind, placebo-controlled trial.** *Lancet.* 2018; **392**(10152): 1036–46.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

42. Nyström L, Larsson LG, Rutqvist LE, *et al.*: **Determination of cause of death among breast cancer cases in the Swedish randomized mammography screening trials. A comparison between official statistics and validation by an endpoint committee.** *Acta Oncol.* 1995; **34**(2): 145–52.
[PubMed Abstract](#) | [Publisher Full Text](#)
43. West THE, Scotland OF, Prevention C, *et al.*: **Computerised record linkage: Compared with traditional patient follow-up methods in clinical trials and illustrated in a prospective epidemiological study. The West of Scotland Coronary Prevention Study Group.** *J Clin Epidemiol.* 1995; **48**(12): 1441–52.
[PubMed Abstract](#) | [Publisher Full Text](#)
44. Brown A, Kirichek O, Balkwill A, *et al.*: **Comparison of dementia recorded in routinely collected hospital admission data in England with dementia recorded in primary care.** *Emerg Themes Epidemiol.* 2016; **13**(1): 11.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
45. Pogue J, Walter SD, Yusuf S: **Evaluating the benefit of event adjudication of cardiovascular outcomes in large simple RCTs.** *Clin Trials.* 2009; **6**(3): 239–51.
[PubMed Abstract](#) | [Publisher Full Text](#)
46. Zhu Y, Matsuyama Y, Ohashi Y, *et al.*: **When to conduct probabilistic linkage vs. deterministic linkage? A simulation study.** *J Biomed Inform.* 2015; **56**: 80–6.
[PubMed Abstract](#) | [Publisher Full Text](#)
47. Tafur Calderon LA, Prada SI: **El sistema general de seguridad social en salud de Colombia.** *Documentos PROESA.* 2019.
48. Ministerio Nacional de Salud y Protección Social: **Sistema nacional de Estudios y Encuestas Poblacionales para Salud: conceptualización y guía metodológica.** Bogotá: Ministerio Nacional de Salud y Protección Social, República de Colombia (Minsalud); 2013.
49. Zea D, Osorio L: **Situación del sistema de vigilancia de casos de dengue en un municipio de Colombia.** *Rev Salud Publica.* 2011; **13**(5): 785–95.
[Reference Source](#)
50. Manrique-Hernández EF, Moreno-Montoya J, Hurtado-Ortiz A, *et al.*: **Performance of the Colombian surveillance system during the COVID-19 pandemic: A rapid evaluation of the first 50 days.** *Biomedica.* 2020; **40**(Supl. 2): 96–103.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
51. Cendales R, Pardo C: **Quality of death certification in Colombia.** *Colomb Med (Cali).* 2018; **49**(1): 121–7.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 04 November 2021

<https://doi.org/10.21956/wellcomeopenres.18817.r46558>

© 2021 Henderson A. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Alasdair Henderson

Department of Non-Communicable Epidemiology, London School of Hygiene and Tropical Medicine, London, UK

In this review the authors aim to demonstrate how routinely collected data can improve randomised clinical trials through enhanced recruitment, expanded follow-up of patients and other advantages. The review is well written, and a compelling case is made for the benefits of routinely collected data. I support the publication of this work, but I have some questions and possible omissions that I think warrant discussion in the review.

Major points:

1. Why does the review focus exclusively on the potential of RCHD to expand and improve randomised clinical trials? Another major advantage of these data is that they have the potential to conduct large observational studies to support epidemiological research (e.g. references 1-3).^{1,2,3} As another major strength of these data, I believe this warrants discussion in this review.
2. Your argument that larger cohorts of patients could reduce random error in these studies is compelling. However, the discussion of possible systemic biases that can be introduced with these data needs to be clarified. If problems such as loss to follow up, selection bias or collider bias are introduced by using RCHD then this will outweigh the benefits of reduced random error.
3. As well as extending and improving randomised trials, could RCHD not be used to replicate RCT designs?^{4,5,6,7} Especially in data with as broad coverage as in Colombia.
4. The review would have been easier to understand if the description of data from Colombia came before the review of improving RCTs with RCHD.
5. The ethical considerations of asking people to potentially be followed for life in routine data needs to be discussed. Will this lead to problems of consent in the collection of data? Does

this introduce selection bias in these data?

Minor points:

1. Either avoid terms such as “a small reduction of 10%” or explain why 10% is deemed a small reduction. Likewise, what do the authors deem as moderate in the following: “Furthermore, moderate random error and biases”?
2. “The randomization of a large number of participants in controlled clinical trials reduces both random error and biases.” There are other considerations in trial design such as blinding and allocation concealment that affect random error and biases regardless of trial size, can these be discussed?
3. Under “Enhancing trial recruitment using routinely collected data”, RCHD may be an efficient method but have possible downsides such as recruiting only those that attend healthcare settings, possibly inaccurate records, and losses to follow up if contact information of individuals changes. This disadvantage should be mentioned alongside the efficiency of RCHD as a recruitment tool.
4. Can terms such as “hard” and “soft” outcomes be defined before being discussed?
5. I would suggest that the “third potential problem” you outline with routinely collected data should be the primary and major potential problem. Also, could you explain in this section why your references and evidence are not from the Colombian data under review? Is this work to be completed?
6. What implication does the delay in data availability (“as long as two years”) have on rules for stopping clinical trials if these data are being used?

References

1. Hernán MA, Alonso A, Logan R, Grodstein F, et al.: Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*. 2008; **19** (6): 766-79 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Bhaskaran K, Douglas I, Forbes H, dos-Santos-Silva I, et al.: Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5·24 million UK adults. *The Lancet*. 2014; **384** (9945): 755-765 [Publisher Full Text](#)
3. Dickerman B, García-Albéniz X, Logan R, Denaxas S, et al.: Emulating a target trial in case-control designs: an application to statins and colorectal cancer. *International Journal of Epidemiology*. 2020; **49** (5): 1637-1646 [Publisher Full Text](#)
4. Hernán M, Robins J: Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available: Table 1. *American Journal of Epidemiology*. 2016; **183** (8): 758-764 [Publisher Full Text](#)
5. Cain L, Saag M, Petersen M, May M, et al.: Using observational data to emulate a randomized trial of dynamic treatment-switching strategies: an application to antiretroviral therapy. *International Journal of Epidemiology*. 2016; **45** (6): 2038-2049 [Publisher Full Text](#)
6. Smeeth L, Cook C, Fombonne E, Heavey L, et al.: MMR vaccination and pervasive developmental disorders: a case-control study. *The Lancet*. 2004; **364** (9438): 963-969 [Publisher Full Text](#)
7. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, et al.: Factors associated with COVID-19-related death using OpenSAFELY. *Nature*. **584** (7821): 430-436 [PubMed Abstract](#) | [Publisher Full Text](#)

Is the topic of the review discussed comprehensively in the context of the current literature?

Partly

Are all factual statements correct and adequately supported by citations?

Partly

Is the review written in accessible language?

Yes

Are the conclusions drawn appropriate in the context of the current research literature?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Epidemiology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 10 August 2021

<https://doi.org/10.21956/wellcomeopenres.18817.r45211>

© 2021 Rojas-Gualdron D. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Diego Fernando Rojas-Gualdron

School of Medicine, CES University, Medellín, Colombia

In this review, the authors present the strengths and weaknesses of using routinely collected health data (RCHD) for the design and conduct of randomized clinical trials. They highlight potential benefits in patient recruitment and representativeness, more extensive follow-up with fewer losses, and cost reduction for large randomized trials. Next, the authors introduce ethical and regulatory considerations regarding access to confidential information and informed consent. Lastly, they proceed to present their perceptions of challenges and opportunities for implementation in Colombia. They conclude that the use of RCHD is feasible in Colombia but requires specific regulation and optimizing registration procedures.

My comments are below:

1. The statement "**Thus, interventions targeting only one or a reduced number of these processes can only be expected to produce small to moderate effects on clinical outcomes.**" requires support by citations. Does this imply that the effect size of an intervention necessarily depends on the number of targeted processes?

2. The statement "**even a small reduction of 10% in CVD mortality could potentially avoid hundreds of thousands of deaths**" requires citation. Additionally, I would not recommend judging a 10% reduction in CVD mortality as a small effect without argumentation.

Two assumptions in the introduction require further argumentation.

1. The assumption that only "**only small to moderate effects can be expected from health interventions**" requires further argumentation. It appears to be based on generic thresholds for effect size. It is not the scope of the review. However, it is relevant to consider if a unified framework of generic effect size criteria (probably brought from experimental research) should be used in the context of outcome and clinical research. I understand that the authors intend to emphasize large samples in randomized studies to overcome methodological issues. However, strong affirmations that are not going to be developed could be avoided.

2. The notion of observational registry-based studies and randomized clinical trials producing opposite results deriving to (apparently) contradictory conclusions may be misleading. This is a core consideration when talking about RCHD. Observational (usually registry-based) and randomized intervention studies do not inform the same facets of the intervention's potential benefit: effectiveness and efficacy. However, they are often misinterpreted. From a methodological point of view, those two designs require different considerations on the implied mechanisms of confounding. I understand that the authors intend to emphasize the need for more randomized studies and the strength of involving RCHD, but a more precise differentiation of this approach from the Real-World evidence framework is required.

3. I find the section "Challenges and opportunities for the use of routinely collected data for trials in Colombia" more descriptive of the Colombian context than reflexive (challenges) or propositive (opportunities). Please consider including more specific recommendations.

For example, the authors state that "Unfortunately, the linkage method is outside the control of researchers since the algorithms used are decided and implemented by the institution that held the databases (NHS)." How to implement statistical linkage techniques in institutions like SISPRO, DANE, or SIVIGILA? Or at least, How likely is that those institutions perform this implementation? Also, the authors state that "misclassification of outcome measures is expected to be non-differential and its impact on the estimations of efficacy is expected to be low as long as the trial is large enough". With so many institutions involved in RCHD this would apply to Colombia? Would a unified framework for outcome definitions (like a Core Outcome Set) be required to optimize RCHD usage in randomized trials?

Those are two examples of aspects presented in the review that the authors could develop further in the section dedicated to the Colombian case. It is not an exhaustive list.

Is the topic of the review discussed comprehensively in the context of the current literature?

Yes

Are all factual statements correct and adequately supported by citations?

Partly

Is the review written in accessible language?

Yes

Are the conclusions drawn appropriate in the context of the current research literature?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Registry-based studies; Outcome research; Evidence synthesis and translation.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.