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

Jonny Alejandro Garcia Luna¹,², Eduardo López-Medina¹,³,⁴, Norman Danilo Maldonado-Vargas⁵, Adrian D. Smith⁶

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

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
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\(^2\). Improved sanitation, hygiene, nutrition, the development of vaccines and the advancement of medical knowledge have been the main drivers of the increased longevity\(^3\). However, millions of premature deaths could be avoided if evidence of better interventions is sought and used to inform medical practice\(^4\). The leading causes of morbidity and mortality worldwide are cardiovascular diseases (CVD) and cancer\(^5\), both of which are multifactorial diseases with complex pathophysiological processes leading to disability and mortality\(^6\). 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 CVD\(^7\); 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 1950\(^8\), while thousands of patients have been randomized to evaluate the efficacy of shorter regimens for this disease\(^9\). Furthermore, the randomization of thousands of patients with corona virus disease 2019 (COVID-19) in the RECOVERY and SOLIDARITY trials was necessary to offer robust evidence of the presence (dexamethasone\(^10\) 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\(^14\). Otherwise, the magnitude of such errors on the estimations of efficacy may obscure or completely reverse the estimated effect of the interventions under evaluation\(^14\). The randomization of a large number of participants in controlled clinical trials reduces both random error and biases\(^15\). Furthermore, both large observational studies and small randomized clinical trials may produce chance or biased results, hindering the implementation of health interventions in clinical practice\(^4\). 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 results\(^18\). 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 diseases\(^20\), 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 supplements\(^22\). 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 strategies\(^23\). 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 trials\(^26\), 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\(^26\),\(^27\).

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 registers\(^28\).
Table 1. Clinical trials using routinely collected healthcare data (RCHD) in their design.

| Trial                     | Country | Sample size | Description                                                                                                                                                                                                 | Use of RCHD                                                                                                                                                                                                 |
|---------------------------|---------|-------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| SAFE-PCI24,28              | 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.                                                  |
| ASCEND12,29–31             | 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). |
| DETO2X32,33               | 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.                                                                         |
| TASTE34                   | 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 mammography35 | Sweden  | >200,000    | Randomized controlled clinical trial of mammography.                                                                                                                                                        | Extended follow up through vital statistics.                                                                                                                                                              |
| AgeX (AgeX)27             | 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 makers. However, people who agree to participate in a clinical trial are different from people who are excluded or do not agree to participate. 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 disease, 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 trials, 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 outcomes. 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 conduction. Nevertheless, RCHD has successfully supplemented outcome information in several clinical trials, reducing loss to follow-up for the most important outcomes. 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 up.

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 interventions. In the Swedish clinical trials of mammography, the average trial duration was 6 years, after which control participants were invited for screenings. 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).

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 vaccine\(^5\).

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 levels\(^6\). 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 bias\(^7\), it also helps to reduce trial costs significantly\(^8,9\). While a typical large randomized clinical trial costs hundreds of millions of US dollars\(^10\), the overall cost of the ASCEND trial was less than 10 million US dollars\(^27\) and the use of the SWEDHEART registry reduced the cost of the TASTE trial by more than 90%\(^11\). 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 issue\(^12\). 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 explanation\(^26\). 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 appointments\(^13\).

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 years\(^14\). 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 participants\(^31\). 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\(^26,34\).

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 cases\(^42\). 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 admissions\(^41\). 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 statistics\(^8\), 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 enough\(^45\).

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 records\(^46\). The two most commonly used methods for data linkage are deterministic and probabilistic\(^47\). 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 databases\(^46\). This method yields the highest specificity\(^46\), 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 matched\(^46\). 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 enough\(^46\).
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 medications\textsuperscript{27}. 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)\textsuperscript{30}. 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 system\textsuperscript{39}. 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\textsuperscript{50} and death certificates\textsuperscript{45}.

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 aspersion areas to estimate the effects of this intervention on the consultation pattern in the affected populations\textsuperscript{40}. 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.
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 THÉ, 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. 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-Calderón LA, Prada S: 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 Nacioanl 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