

A new scope and a vision for record linkage studies

GIOVANNI CORRAO⁽¹⁾, CARLO LA VECCHIA^(2, 3)

(1) *Giovanni Corrao, Department of Statistics and Quantitative Methods, Division of Biostatistics, Epidemiology and Public Health, University of Milano-Bicocca, Milan, Italy*

(2) *Department of Epidemiology, IRCCS-Istituto di Ricerche Farmacologiche "Mario Negri", Milan, Italy*

(3) *Department of Clinical Sciences and Community Health, Unit of Medical Statistics and Biometrics, University of Milano, Milan, Italy*

CORRESPONDING AUTHOR: *Giovanni Corrao, Department of Statistics and Quantitative Methods, Division of Biostatistics, Epidemiology and Public Health, University of Milano-Bicocca, Via Bicocca degli Arcimboldi, 8, Building U7, 20126 Milan, Italy. Tel: +39 02 64485854.*

Fax: +39 02 64485899. e-mail: giovanni.corrao@unimib.it

DOI: 10.2427/9036

Record linkage - i.e. the use and linkage of electronic, mainly administrative databases - has been utilized in epidemiologic research for over three decades now, starting from the system implemented in a few US health maintenance organizations (HMO, i.e. the Kaiser Permanente network), a few Nordic countries systems, or the UK general practice research database (UK GPRD, now clinical practice research datalink, CPRD) [1, 2].

Record linkage-based studies were used essentially to investigate the implications and side effects of drugs and, more in general, medical interventions, but allowed often broader etiologic inference [3-5]. Thus, the absence of excess breast cancer risk following (induced) abortion was established in a Danish record linkage study, since studies based on ad hoc data collection (cohort and, mostly, case-control studies) were affected by information bias [6].

The progresses in computing and epidemiologic methodology allow now important developments in the use of record linkage-based research. This issue of EBPH includes therefore a thematic series of papers on the scope, methods and perspectives of the use of electronic databases in epidemiologic research.

The paper by Romio et al. [7] describes why and how health care/policy decision making can benefit from critical and research oriented use of administrative data. It describes methods to investigate real-world clinical practice, lists potentialities of routinely collected data, reviews their global availability, and outlines open challenges in this field.

The paper by Corrao [2] is a key lecture for anybody interested in methods and applications of routinely collected and electronically stored information on healthcare utilisation. It addresses several methodological issues, starting from strategies of sampling within a large cohort, as an alternative to analysing the full cohort, to methods for controlling outcome and exposure misclassification. It also considers techniques taking into account both measured and unmeasured confounders, as well as the issue of random uncertainty in the framework of studies using healthcare utilisation data. It thus provides readers with a methodological framework, while commenting on the value of new techniques for more advanced users.

The paper by Patorno et al. [8] provides an overview of the use of propensity score methodology for pharmacoepidemiology research with large healthcare utilization databases. In particular, it addresses open issues on covariate selection, the role of automated techniques for addressing unmeasurable confounding via proxies, strategies to maximize clinical equipoise at baseline, and the potential of machine-learning algorithms for optimized propensity score estimation.

The paper by Stendardo et al. [9] addresses a number of legal issues, starting from how to guarantee data confidentiality and patient anonymity, since individual information is often accessed without the

patient concerned being involved. It also considers other relevant aspects, such as those connected with the nature of sensitive data, data ownership, and applicable penalties in the event of a breach of the relevant legal provisions.

In this thematic issue, there are two papers providing innovative examples of the utilization of administrative data in medical research. The paper by Negri et al. [10] investigates the use of 17 new targeted high cost drugs in Lombardy oncology practice between 2006 and 2010 using data from electronic health-care databases, with the major aims to provide estimates of the incidence of serious adverse events of novel anticancer therapies, and of overall and disease free survival in clinical practice, overall and in selected subgroups [11].

The paper by Zins et al. [12] illustrates the CONSTANCES cohort, designed as a randomly selected representative sample of French adults aged 18-69 years at inception; it will include 200 000 subjects over a five-year period. At inclusion, the selected subjects are invited to fill a questionnaire and to attend a Health Screening Center (HSC) for a comprehensive health examination: weight, height, blood pressure, electrocardiogram, vision, auditory, spirometry, and biological parameters; for those aged 45 years and older, a specific work-up of functional, physical, and cognitive capacities is performed. A biobank will be set up. The follow-up includes a yearly self-administered questionnaire, and a periodic visit to an HSC. Social and work-related events and health data are collected from the French national retirement, health and death databases, thus providing an example of the wide scope of new record linkage studies.

The last paper by Corrao et al addresses the main aspects of the Carry out a Repository for Administrative and Clinical data Knotting (CRACK) program [13]. This wide-range project is aimed at implementing a repository through the integration of different data sources for addressing questions still open in the etiology and management of common and clinically relevant diseases. The questions addressed include disease frequency and prognosis (prevalence, incidence, survival), healthcare utilization patterns and safety, effectiveness and cost-effectiveness profiles in real-world clinical practice.

Methodological issues as well as specific clinical and public health questions will be investigated by exploiting them from the decision maker as well as the researcher point of view. Thus, the CRACK program will implement a “flywheel” allowing health authorities to obtain solid data and strong evidence for addressing health policies, as well as researchers to address innovative questions, produce and publish findings of global interest.

This series of papers provides therefore a comprehensive and up to date vision for the status and development of record linkage studies.

ACKNOWLEDGEMENTS: *the work of Carlo La Vecchia was partly supported by the Italian Association for Cancer Research (AIRC- project no. 10068).*

References

- [1] Schneeweiss S, Seeger JD, Jackson JW, Smith SR. Methods for comparative effectiveness research/patient-centered outcomes research: from efficacy to effectiveness. *J Clin Epidemiol* 2013; 66(8 Suppl): S1-4
- [2] Corrao G. Building reliable evidence from real-world data: methods, cautiousness and recommendations. *Epidem Biostat Public Health* 2013; 10(3). DOI: 10.2427/8981
- [3] Corrao G, Zambon A, Nicotra F, et al. Persistence with oral and transdermal hormone replacement therapy and hospitalisation for cardiovascular outcomes. *Maturitas* 2007; 57(3): 315-24
- [4] Corrao G, Zambon A, Conti V, et al. Menopause hormone replacement therapy and cancer risk: an Italian record linkage investigation. *Ann Oncol* 2008; 19(1): 150-5
- [5] Corrao G, Parodi A, Nicotra F, et al. Better compliance to antihypertensive medications reduces cardiovascular risk. *J Hypertens* 2011; 29(3): 610-8
- [6] Beral V, Bull D, Doll R, et al. Breast cancer and abortion: collaborative reanalysis of data from 53 epidemiological studies, including 83000 women with breast cancer from 16 countries. *Lancet* 2004; 363(9414): 1007-16
- [7] Romio S, Sturkenboom M, Corrao G. Real-world data from the health decision maker perspective. What are we talking about? *Epidem Biostat Public Health* 2013; 10(3). DOI: 10.2427/8979

- [8] Patorno E, Grotta A, Bellocco R, Schneeweiss S. Propensity Score Methodology for Confounding Control in Health Care Utilization Databases. *Epidem Biostat Public Health* 2013; 10(3). DOI: 10.2427/8940
- [9] Stendardo A, Preite F, Gesuita R, et al, SOS working group. Legal aspects regarding the use and integration of electronic medical records for epidemiological purposes. *Epidem Biostat Public Health* 2013; 10(3). DOI: 10.2427/8971
- [10] Negri E, Rossi M, Bonifazi M, et al. Clinical use, safety and effectiveness of novel high cost anticancer therapies after marketing approval: a record linkage study. *Epidem Biostat Public Health* 2013; 10(3). DOI: 10.2427/9016
- [11] Bonifazi M, Rossi M, Moja L, et al. Bevacizumab in clinical practice: prescribing appropriateness relative to national indications and safety. *Oncologist* 2012; 17(1): 117-24
- [12] Zins M, Berkman LF, Goldberg M, and the CONSTANCES team. The CONSTANCES cohort, an epidemiological research infrastructure. Methods and results of the pilot phase. *Epidem Biostat Public Health* 2013; 10(3).DOI: 10.2427/8921
- [13] Corrao G, Cesana G, La Vecchia C, et al. The CRACK program: a scientific alliance for bridging health care research and public health policies in Italy. *Epidem Biostat Public Health* 2013; 10(3).DOI: 10.2427/8980

