INTRODUCTION

HEALTHCARE ADMINISTRATIVE DATABASES AND ALGORITHMS: AN ONGOING EPIDEMIOLOGICAL HISTORY

Lorenzo Simonato,[1] Giovanni Corrao[2]
[1] Department of Cardio-Thoraco-Vascular Sciences and Public Health, University of Padua, Padua (Italy)
[2] Università degli Studi di Milano-Bicocca, Milan (Italy)
Corresponding author: Lorenzo Simonato; lorenzo.simonato@unipd.it

This supplement of *Epidemiologia&Prevenzione* is an extension and update of the Supplement published in 2008 on *E&P*[1] and reports the results of the research activities of a Collaborative Working Group assembled with the goal of systematically collecting and reviewing the available published evidence on the use of algorithms in case identification in Italy during the past ten years.

The availability and use of electronic health data has grown during the last twenty years, and Italy was a pioneer in this sector since the use of electronic health database started in early Eighties'.[2,3]

The first evidence in diseases registration systems is associated with modernization of cancer registration techniques.[4,5] For the first-time, administrative health electronic archives (AHEA) were used with the goal of providing population-based frequency estimates of a chronic disease like cancer. On this occasion, a cancer-specific algorithm was developed, based on the combined use of death certificates, hospital admissions, and pathology records.[6,7]

Subsequently, an effort was made by a national collaborative group to try to extend this methodological approach to other chronic diseases.[1]

We refer here and in the other parts of this Supplement to health archives generated within the activities of the National Health System (NHS) and, as such, subject to the rules of the laws on confidentiality. The information directly or indirectly available from social networks or other commercial systems, so called Big Data,[8] are of different nature and significance and was not considered by the Working Group.

During the last ten years, the availability and use of electronic health data has grown mainly in three directions:

- Punctual estimates of disease frequencies in a given population and at a given point in time;
- Measuring the performance of various functions of the NHS;
- Building up Integrated Epidemiological Systems (IES) for the follow-up of the general population over time, with the aim of understanding the dynamic health characteristics of the population, monitoring appropriateness, safety, effectiveness, accessibility, and equity healthcare pathways[9], and developing the appropriate actions.

All these different approaches are not alternative or conflicting, but just different pathways to achieve as great a knowledge as possible of the real, underlying health characteristics, which might be defined as "symptoms", in a given population. Yet, in a more advanced methodological perspective, they could rather be considered as part of a unique system in which the population file is the matrix and the activities of the NHS are a kind of machine registering all the health-related events occurring among the individual members of the population. If we consider these elements as a whole, clearly the organization of the architecture of the system comes first, while the algorithms are more dependent on the various specific analytical activities to be performed within the system. In figure 1, a conceptual framework is tentatively sketched.

Inevitably, no system is perfect, as any system relies on the availability and quality of data input, which can be quite variable from population to population. Still, there

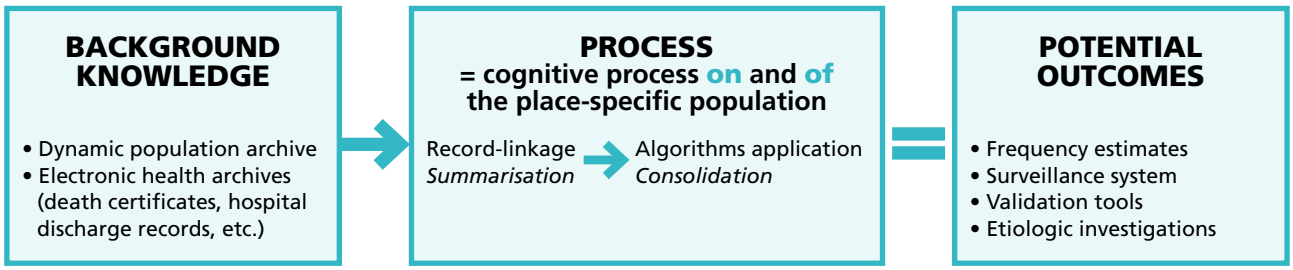

Figure 1. Theoretical frame of data processing.

are two important advantages: **1** the potential coverage of the whole population rather than selected parts or samples of it, thus avoiding selection bias or statistical power limitations; **2** the possibility of introducing in the system the concept and practice of life-time dimension.

Technically, more or less since the turn of the century, it has been feasible in Italy to follow an individual from birth to death through the archives of the NHS. This situation recalls the hypothesis foreseen in 1946 by the chief of the U.S. National Office of Vital Statistics, Halbert Dunn, who anticipated the possibility, at the time only theoretical, of building up a "book of life" composed of all the recorded events occurring to each single member of the community from birth to death using record-linkage technology.[10]

Life Long Studies (LLS) are of increasing importance in the field of public health,[11] as these individual "trajectories", traced through health monitoring systems embedded in the NHS, can be organized epidemiologically in population-based prospective studies for which an unlimited follow-up can regularly be programmed and never discontinued as long as the NHS exists. If we consider LLS as the most informative approach to population knowledge and consequently to actions aiming at improving the health conditions of a population or community, we should give priority to the two most crucial steps of the trajectories occurring within the population members: the beginning of life and the period preceding death.

Firstly, the beginning of life is rich in information registered by the NHS (CEDAP files, hospital admissions, first aid departments, etc.) which potentially permits the setting up of birth-cohorts[12] and the definition of subgroups at high risk requiring targeted assistance.[13] It is in fact plausible that an important component of adult health status originates from early and very early life, with important implications for long-term preventive action.

Secondly, the time preceding the end of life has an increasing relevance in our Country and other Western Countries in which life expectancy has approximately doubled during the last 150 years. The implications of this astonishing phenomenon are medical, socio-economical and ethical, and strongly challenge, besides the efficiency and efficacy of the National Health Systems, the basis of our societal organization. The possibility of dynamically observing and interpreting in detail within the general population what happens during the last period before the end of life can open new perspectives of better understanding physiological aspects of the aging process and revising some basic concepts traditionally associated with the relationship between life and death.

However, barring any relevant changes in the data flows of the NHS in the future, complete LLS will only be available in about 60-70 years, as data flows in all Italian regions (although still with variable completeness and quality and difficult to verify while respecting the current privacy rules) have been available only since 2010. In the meantime, electronically stored healthcare data is available in all Italian regions, making it possible to carry out population-based monitoring systems aimed to verify and compare appropriateness, effectiveness, safety, costs, and accessibility of the NHS, and services guaranteed through the Essential Assistance Levels (LEA).[14] All this, however, requires a huge, coordinated methodological effort to ensure the quality standards of the evidence generated using electronic health records. Among these, the algorithms for case identification play a necessary – although still insufficient – role.

We should, at this point, also stress some ethical aspects that cannot be separated from the existence of the NHS. National health systems, which started to be formally introduced in a number of European Countries after World War II, should not simply be considered an effort to rationalize the medical assistance to face the new challenges of medical science in the second half of the 20th century. The conceptual background of NHS is based on a set of important principles: first, the principle of solidarity amongst the members of a community considering their intrinsic vulnerability as human beings. The fear of losing our wellness is shared with the other members of the community in an organized structured network and, as written in the Bellagio Statement of Principles: "public health efforts are more likely to succeed in an atmosphere of solidarity and trust".[15] In turn, differential levels

of solidarity are strongly associated with degrees of equality, community membership, and relational liberty.[16]
With respect to the theme of this monograph, solidarity also consists in implicitly making available our health data (for example, medical prescriptions, hospital diagnosis, outpatient visits) provided that: • confidentiality is guaranteed; • the data are useful for generating evidence used to support decision makers in the process of continuous improvement of medical services (i.e., evidence from the past to better cure patients in the future). Rightly, the focus on the first constraint is maximum. The recent entry into force of the European regulation on the protection of individuals with regard to the processing and circulation of personal data (Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC), and the publication of the adaptation of national legislation to the European regulation (Legislative Decree 10 August 2018, No. 101, published in the GU General Series No. 210 of 4 September 2018), strive to offer ample guarantees on this issue. The legal and political sensitivity on the second point is less felt. It is not enough to have data: we should organize them and treat them according to rules of good practice to generate credible evidence from secondary data. We hope that the Ministry of Health, with the technical support of the Istituto Superiore della Sanità (National Health Institute, NHI) , scientific societies and universities, will promote a working group to proceed in this direction.
In the meantime, the mixed academic-NHS working group that produced this monographic issue of E & P has tried to give a contribution to the knowledge and standardization of the algorithms for case finding. The advantages of using comparable algorithms in case definition should be rather evident in an epidemiological context and their regular and standard application within the activities of health surveillance systems should be recommended.
There is still a long way to go, as for many diseases we do not have consolidated algorithms. Results from proper validation processes are rarely available, and our knowledge of diseases in a rapidly ageing population is still based on past classic evidence. Another weakness of these integrated systems concerns the difficulty in disseminating and sharing with community members correct information on their products. The evidence available generated by knowledge of health characteristics of the populations under follow-up is often ignored or misinterpreted. If a population is the matrix of the cognitive process, that population ought to be not only correctly informed of its health status, but also placed in the condition of sharing the responsibility of understanding and making decisions on its health prospects.

REFERENCES

1. Simonato L, Baldi I, Balzi D, et al. Obiettivi, strumenti e metodi per un utilizzo epidemiologico di archivi sanitari elettronici correnti in diverse aree italiane. Epidemiol Prev 2008;32(3) Suppl:5-14.
2. Maggini M, Salmaso S, Spila Legiani S, Caffari B, Raschetti R. Epidemiological use of drug prescriptions as markers of disease frequency: An Italian experience. J Clin Epidemiol 1991;44(12):1299-307.
3. Costa G, De Maria M, Bisanti L, et al. Uso dei dati amministrativi per la ricerca epidemiologica. Epidemiol Prev 1988;35:40-46.
4. Simonato L, Zambon P, Rodella S, et al. A computerised cancer registration network in the Veneto region, north-east of Italy: a pilot study. Br J Cancer 1996;73(11):1436-39.
5. Black RJ (ed). Automated Data Collection in Cancer Registration. Lyon: IARC; 1998.
6. European Network of Cancer Registries, Tyczyński JE, Démaret E, et al. Standards and Guidelines for Cancer Registration in Europe: The ENCR Recommendations : Volume I. Lyon: IARC; 2003.
7. Eastman P. Bringing Cancer Registries into the 21st Century: A Tale of Three Countries. Oncol Times 2006;28(15):6-9.
8. Weber GM, Mandl KD, Kohane IS. Finding the Missing Link for Big Biomedical Data. JAMA 2014;311(24):2479-80.
9. Corrao G, Mancia G. Generating Evidence From Computerized Healthcare Utilization Databases. Hypertension 2015;65(3):490-98.
10. Dunn HL. Record Linkage. Am J Public Health Nations Health 1946;(36):1412-16.
11. Richter M, Blane D. The life course: challenges and opportunities for public health research. Int J Public Health 2013;58(1):1-2.
12. Canova C, Pitter G, Schifano P. A systematic review of epidemiological cohort studies based on the Italian Medical Birth Register. Is it time to think of a multicentric birth cohort? Epidemiol Prev 2016;40(6):439-52.
13. Cantarutti A, Franchi M, Monzio Compagnoni M, Merlino L, Corrao G. Mother's education and the risk of several neonatal outcomes: an evidence from an Italian population-based study. BMC Pregnancy Childbirth 2017;17(1).
14. Corrao G, Rea F, Martino MD, et al. Effectiveness of adherence to recommended clinical examinations of diabetic patients in preventing diabetes-related hospitalizations. Int J Qual Health Care 2018. doi: 10.1093/intqhc/mzy186
15. Bellagio Group. Bellagio Statement of Principles. 2007. Available from: http://www.bioethicsinstitute.org/research/global-bioethics/flu-pandemic-the-bellagio-meeting (last accessed: 7 September 20149.
16. Jennings B. Relational Liberty Revisited: Membership, Solidarity and a Public Health Ethics of Place. Public Health Ethics 2015;8(1):7-17.