

# Methodological perspectives for register-based data analysis

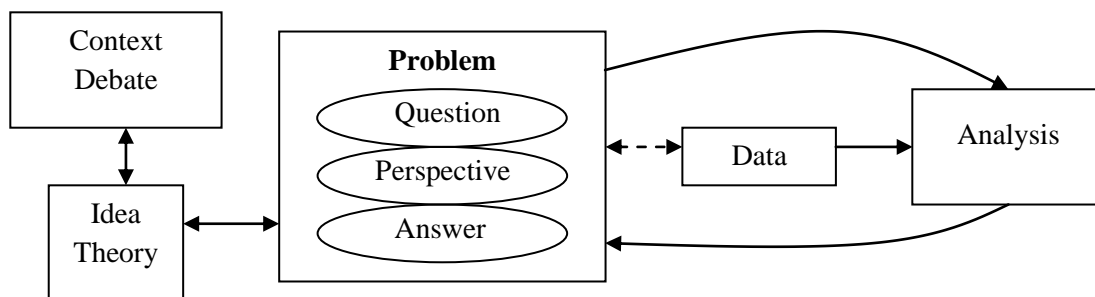
Reijo Sund

National Institute for Health and Welfare, Service Systems Research Unit

[reijo.sund@thl.fi](mailto:reijo.sund@thl.fi)

Data have been produced for hundreds of years. Advances in information technology have made it possible to produce and store all kinds of data effectively. Administrative registries have been at the forefront of data gathering, with a growing demand for evidence-based information to support decision-making and for other administrative purposes.

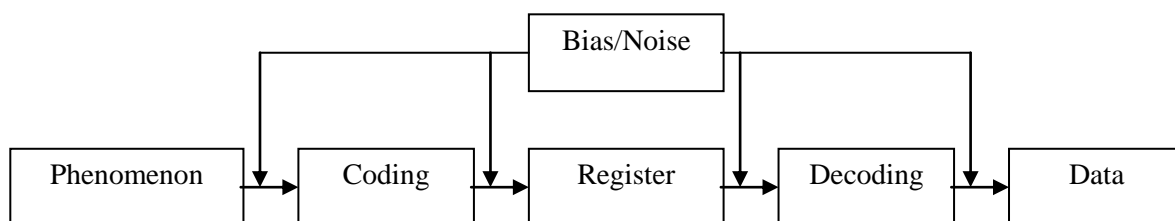
The general aim in the utilization of administrative registers for production of statistics or for research purposes is to transform the raw data in the register into useful information. Unfortunately, as long as there have been administrative registers, there has been a bottleneck in this kind of transformation. The main problem with administrative data is that the measurement can only be based on this existing secondary data, originally produced for some other purposes than the current utilization task at hand.



**Figure 1. Research process schema**

It has been suggested that reasonable results can be achieved by considering the whole series of actions required to transform data into information as a knowledge discovery process. In this sense, it is convenient to consider data analysis as a collection of tasks during the research process. A description of the phases commonly encountered during an empirical research process that incorporates register-based data is presented in Figure 1. The main phases in such process are: understanding the phenomenon, understanding the problem, understanding data, data preprocessing, modeling, evaluation and reporting. The most important difference in this kind of process in comparison to a standard scientific inquiry is the connection between the problem and the data, i.e. the challenge of secondary data.

This challenge can be illustrated in terms of information communication (Figure 2). First, it is assumed that some phenomenon exists that can be observed. Since it is impossible to completely observe all details or perform exact measurements, some kind of coding is used to describe things. This coded signal is then stored in a database. The noise and bias can be interpreted as an explanation for measurement compromises, possible inconsistencies and coding errors, and coding practices existing in the stored signal. When this signal is then utilized, it must be decoded into a suitable form. This phase is also subject to noise and bias caused by incompatibility of choices and interpretations made by the data producer and the data user. Even the decoded signal (data) is not a final phase in the research process, because further analysis and processing is needed in order to transform the data into information. Even though this is a very simple and technical representation of communication, it seems to contain the essential elements needed in the common-sense understanding of secondary data.



**Figure 2. Schematic diagram of information communication via administrative registers**

The schemas in Figures 1 and 2 give a basic overview of the process of register-based data analysis, but leave many pragmatic details unanswered. Even though the actual realization of the process is determined by the research problem and the available register data, the effective use of register-data presumes skills in at least four areas: in the principles of measurement, in information science, in statistics and in the theory of the subject matter.

By explicitly considering aspects of theory, observation, measurement, operationalization, data, and data sensitive interpretation while trying to transform data into information helps to spot the problematic issues in data or in theory:

- In order to find some shared perspective between the original and intended data utilization purposes, there is a need for a conceptual representation of each object of interest in the terms of knowledge, logical, and data components.
- There are two main categories of concept-data relationships: 1) the stable one where the need for additional background information is minimal; and 2) the abstracted one where the final data are a result of some intelligent transformation of available data based on the cognitive fit between the theoretically suitable and real observables.
- A sophisticated preprocessing—which is full of ideologically dependent qualitative choices—in order to scale matters down to a size more suitable for specific analyses is the most important and time-consuming part of register-based data analysis.

In this presentation these ideas and approaches will be briefly reviewed with examples in order to demonstrate a methodological framework that may help to analyze register-based data more effectively.

### **For more information**

Sund, Reijo (2003): Utilisation of Administrative Registers Using Scientific Knowledge Discovery. *Intelligent Data Analysis* 7:6, 501-519.

Sund, Reijo & Nylander, Olli & Palonen, Tuula (2004): Raa'asta rekisteriaineistosta terveystieteellisesti relevanttiin informaatioon. *Yhteiskuntapolitiikka* 69:4, 372-379. <http://yp.stakes.fi/FI/arkisto/sisallys/2004/2004.htm#4>

Sund, Reijo: Utilization of routinely collected administrative data in monitoring of aging-dependent hip fracture incidence (2007): *Epidemiologic Perspectives & Innovations* 4:2. <http://dx.doi.org/10.1186/1742-5573-4-2>

Sund, Reijo (2008): Methodological Perspectives for Register-Based Health System Performance Assessment. Developing a Hip Fracture Monitoring System in Finland. Stakes Research Report 174. National Research and Development Centre for Welfare and Health, Helsinki. <http://urn.fi/URN:ISBN:978-951-33-2132-1>

Sund, Reijo (2010): A framework for evaluating the quality of administrative data for research purposes. Proceedings of Q2010 European Conference on Quality in Official Statistics. Available at [http://q2010.stat.fi/media/presentations/sund\\_q2010\\_paper.pdf](http://q2010.stat.fi/media/presentations/sund_q2010_paper.pdf)