Validation of healthcare databases: why to do it and how to do it

Aldana Rosso
Epidemiology and Register Centre South, Skåne University Hospital, Lund, Sweden
Validation of healthcare databases: why to do it and how to do it

Aldana Rosso

Abstract.
Healthcare databases are unique tools in public health research to evaluate the healthcare performance and treatment compliance. Knowledge about the data quality is required to enable interpretation of the results from statistical analyses. We summarize different sampling methods suitable for the validation of healthcare databases and discuss their advantages and disadvantages.

Key words: validation, healthcare databases, national registries, register-based research, statistics, sampling theory.

1. Introduction

It is of paramount importance to follow up and evaluate how the current healthcare policies impact the health of the population. A way to monitor current policies is to analyse data collected in healthcare databases and quality registries. Quality registries are healthcare databases that have been developed as a complement to medical charts to collect variables that are not systematically recorded. Quality registries focus on a specific disease group and collect information at national, regional or local level. One of the main advantages of quality registries is that they facilitate the comparison of interventions across different units, since the same variable definitions are used. For example, in Sweden there are approximately 100 national quality registries that get financial support from the national government.

When performing both qualitative and quantitative analyses using healthcare databases it is assumed that the data is complete and correct. If data is missing or incorrect this may affect the quality of the analyses, and therefore wrong conclusions may be reached. If the incorrect and missing data are completely random, meaning that they are not related to the outcome or to other variables, the power of the analysis will be lower (compared to if all data were correct and no data missing) but the results would be unbiased. In this case, more observations will be needed to achieve the same power in the statistical analyses. On the other hand, if the data is missing or incorrect due to other reasons, this may introduce a systematic bias and
wrong conclusions would be reached. A clear example of how missing data may bias the results is presented in a study performed by Ferreira-González et al. (Ferreira-González et al., 2009). The authors compare, among other variables, the mortality risk of patients registered in the Spanish Acute Coronary Syndrome Registry with the mortality of patients not included (missing) in the registry. The authors conclude that the mortality risk is almost three times higher in the missing population (Ferreira-González et al., 2009). An example of how relative small errors may affect certain types of statistical analysis is presented by Ranstam et al. using data from the Swedish Knee Arthroplasty Registry (Ranstam et al., 2008). Briefly, the authors calculated how the revision risk and the ranking of number of revision in hospitals is affect by missing and incomplete data. The authors show that whereas minor registration incompleteness has little effect on the revision risk, it can lead to major errors in the ranking of hospitals (Ranstam et al., 2008). More recently, a validation study performed on the Dutch National Intensive Care Evaluation (NICE) Registry found that the registry had approximately 1% incorrect values for in-hospital mortality (Koetsier et al., 2013). However, this was sufficient to affect the ranking of one of the participating intensive care unit. Before the correction of the errors this unit was performing better than other units and after the correction the performance became average. These examples illustrate that to draw correct conclusions from the statistical analyses we need to know the mechanisms behind the errors and missing data. In order to obtain information about why errors in the database appeared we do not need to examine every point in the database. By performing a validation study, the same information can be obtained with significantly less effort.

Sampling theory is a field within statistics that focuses on how to choose different elements from a population to obtain a representative sample (Paul Levy and Stanley Lemeshow, 2008). In order to estimate the proportion of errors in a database, we first select a sample using sampling theory, so we assure that the sample is representative for the entire database. Secondly, the errors in this sample are calculated and finally the results are extrapolated to the entire population. A validation study may focus on different quality aspects, for example, are the patients that we want to study included in the database? How many incorrect registrations are in the database? In this article, we will discuss some statistical methods that are useful to estimate the total number of errors in healthcare databases.

2. Statistical Methods

Errors in a healthcare database are due to different causes and typically difficult to detect. For example wrong information might be registered in the patient’s health records, and this information is then stored in the database. If the database is filled manually, the data source may be correct but wrong data may be registered due to data entry mistakes. Finally, if the database automatically collects data from different sources, errors in the programming may cause that the wrong data is loaded into the database.

Errors occurring in the data source are very difficult to detect, since we usually assume that the data source contain the “true values”. In order to investigate the quality of the data source (in most cases patients’ health records), a random sample
of the database can be selected and then an expert group goes through the cases. The quality of the source is then assessed by the expert group. This process is called adjudication. The results obtained for those cases are then extrapolated to the entire database.

In order to obtain information about the errors that happen during and after registration, a similar procedure is performed. In this case the database is compared to the patients’ health records, which are assumed to be correct. When the information in the database is concordant with the one in the patients’ health records, the data is defined as correct; otherwise the data is defined as wrong, even if the original health records are not correct.

The first step in the design of a validation study is to obtain an estimation of the proportion of errors that are present in the database. Unfortunately, this information is typically unknown. An earlier study from the Dutch National Intensive Care Evaluation (NICE) registry indicated that the proportion of errors may be approximately 5% and the proportion of incomplete data about 3% (Arts et al., 2002). An audit performed on the UK Cardiac Surgery database showed that approximately 25% of essential data elements were missing in the registry (Fine et al., 2003). We also need to decide the precision of the estimate of the proportion of error that we want to obtain. For example, we could assume that the proportion of errors in the database is 5% and the largest confidence interval for the error estimate that we could accept is up to 7%, this being the upper limit of the confidence interval. Furthermore, in those projects where the database will be compared to the patients’ original health records, practical information about how these records are going to be obtained is required. For example: do we need to visit several clinics in order to obtain the original health records? Is it feasible to involve all clinics present in the database? All this information is considered in order to choose the most suitable design for the validation study.

There are many ways to draw a representative sample from the database for validation purposes. The simplest strategy is called simple random sampling. Briefly, some elements of the database are selected randomly and then this data is compared to the patients’ health records (see Figure 1 A). Some of the disadvantages are that, since it is a random sample, we cannot assure in advance that all groups of patients are represented in the validation sample. Sometimes we may have some auxiliary information that could help to improve the sampling design, for example if we know that some clinics have more errors than others. In this case it could be more efficient to choose another sampling strategy that takes this factor into account.

In order to incorporate auxiliary information about the registry, more advanced designs are needed. These are typically obtained by combining simple random sampling with cluster sampling and stratification, which is called multistage sampling. In some cases, it could be an advantage to first stratify (divide) the database into non-overlapping subpopulations (strata) and then obtain a random sample for each subpopulation (see Figure 1 B). For example, several clinics contribute with data to a healthcare registry. The participating clinics have different routines for registration, and these differences may affect the data quality. Since we cannot assume a priori that the clinics have similar error levels we would like to
obtain a separate error estimate for each clinic. Therefore, stratification by clinic is a more suitable alternative compared to simple random sampling. The overall error rate is then obtained by pooling the individual estimates. The main advantage of stratification is that error estimations are obtained for each level of the stratification variables. However, this also implies that we need to compare records from the database for all stratification levels. In the previous example, this means that all the clinics need to participate in the validation study. If the variation of the amount of incorrect data between the clinics is large, this design is more efficient than simple randomization (Risto Lehtonen and Kari Djerf, 2008). If the proportion of error is similar for all stratification levels, this method does not require a larger validation sample than simple randomization.

Cluster sampling is a suitable sampling method in those cases where the resources are limited just to validate a certain number of levels of a variable (for example clinics) since only a group of levels is used in the validation (see Figure 1C). The selection of the levels can be done by giving more weight to large levels (larger clinics) or by selection randomly a number of levels. Contrary to stratification, this sampling method is optimal when the error frequency is similar for the different levels. The main disadvantage of this method is that it usually requires a larger sample size than simple randomization (Risto Lehtonen and Kari Djerf, 2008).

It is also possible to combine cluster sampling and stratification. For example, we could first select some clinics (clusters) by giving more weight to larger clinics. Then, within the selected clinics, we could stratify by diagnosis and finally randomly select a group of observations. Alternatively one could stratify by region or province, then select some clinics within each region, then stratify by diagnosis and finally randomly select the same number of observations for each diagnosis category.

### 3. Simulation study

Since all databases are different it is not possible to recommend a general validation approach that is optimal for all cases. In order to decide which validation strategy is most suitable for a certain project Monte Carlo simulations can be performed. Different scenarios are programmed and the size of the validation sample is
calculated. The simulations for the examples discussed above are rather simple to perform and a lot of resources can be spared. Standard statistical packages such as Stata, SAS, and R have built in routines to perform these simulations.

To illustrate some common sampling strategies we present simulations for three different cases. Let’s suppose that we would like to validate a healthcare registry with 2,500 records. The registry has three participating clinics with 33 % of the records each. From previous studies, we would expect the percentage of incorrect data to be around 2 %. We are willing to accept that the estimate of the percentage of error to be up to 4 %. All the simulations were performed using Stata/IC version 13.1 (StataCorp LP, Texas, USA).

3.1 Simple random sampling

If we can assume that the clinics have similar percentage of errors then simple random sampling is a suitable strategy. An additional advantage is that this technique is easy to explain to the registry user community. Under the conditions named above, the validation sample size would be about 300 records.

3.2 Stratification at clinic level

Stratification at clinic level is optimal if we need accurate estimates for the amount of errors for each clinic. In this case, we draw a random sample of the patients for each clinic and compare the data with the original records. Under the conditions named above, by sampling 15 % of the patients per clinic (375 records in total) we would obtain an error estimate around 3 %.

3.3 Cluster sampling at clinic level

Let’s assume that due to lack of resources we can only sample two clinics and 50 patients each. Cluster sampling is a strategy that allow us to accommodate these constrains in the validation design. Under this conditions, we would obtain an error estimate just below 6 %. In order to obtain an error estimate around 3 % we would need to sample 440 patients in total.

4. Conclusion

In conclusion, the results of the analyses performed using healthcare databases are reliable only if we know the amount of correct data in the database and the mechanisms behind the errors and missing values. The verification of every single data point is not needed in order to obtain information about the errors in the database. By performing a validation study optimized to suit a certain database, a random sample is drawn and analysed. The results are then extrapolated to the entire database. Validation studies are very flexible and provide a simple way to obtain information about how the quality of the database affects our conclusions.

Acknowledgments
The author would like to Prof. Ulf Stenevi at the Department of Ophthalmology, Sahlgren's University Hospital, Mölndal, Sweden for interesting discussions about how to validate data from national quality registries.

5. References


