Proceedings of Statistics Canada Symposium 2016
Growth in Statistical Information: Challenges and Benefits

The challenges of linking and using administrative data from different sources

Philippe Gamache[1]

Abstract

At the Institut national de santé publique du Québec, the Quebec Integrated Chronic Disease Surveillance System (QICDSS) has been used daily for approximately four years. The benefits of this system are numerous for measuring the extent of diseases more accurately, evaluating the use of health services properly and identifying certain groups at risk. However, in the past months, various problems have arisen that have required a great deal of careful thought. The problems have affected various areas of activity, such as data linkage, data quality, coordinating multiple users and meeting legal obligations. The purpose of this presentation is to describe the main challenges associated with using QICDSS data and to present some possible solutions. In particular, this presentation discusses the processing of five data sources that not only come from five different sources, but also are not mainly used for chronic disease surveillance. The varying quality of the data, both across files and within a given file, will also be discussed. Certain situations associated with the simultaneous use of the system by multiple users will also be examined. Examples will be given of analyses of large data sets that have caused problems. As well, a few challenges involving disclosure and the fulfillment of legal agreements will be briefly discussed.

Key words: administrative data; linkage; quality; disclosure; monitoring.

1. Quebec Integrated Chronic Disease Surveillance System (QICDSS)

1.1 Introduction

For more than 15 years, the Institut national de santé publique du Québec (INSPQ) has been monitoring chronic diseases under its mandate from the Ministère de la Santé et des Services sociaux du Québec (MSSS). Initially, the population was monitored using administrative files (death, hospitalizations) and surveys, but the data sources were processed independently. For the past few years, linking a number of administrative files has greatly improved the reliability of estimates of many indicators produced by the INSPQ, and particularly with respect to chronic disease surveillance.

The system based on this recent linkage of data from administrative files is known as the Quebec Integrated Chronic Disease Surveillance System (QICDSS). To date, QICDSS covers the period from April 1996 to March 2014. While the system offers numerous benefits, it is also the source of many methodological, technological and technical challenges. How should five databases, not only administered by different bodies but for which the primary use is not monitoring chronic diseases, be properly linked? Is it possible to reduce the impact of varying data quality? Can multiple statisticians, analysts, students and interns use QICDSS simultaneously without creating problems? What strategies should be used to facilitate the system's use while respecting all legal requirements?

1.2 The differences between the QICDSS's five databases

Five main administrative files are linked to make up the QICDSS. They are the death file, the hospitalization file (Maintenance et exploitation des données pour l'étude de la clientèle hospitalière), the register of persons insured under Quebec's health insurance (FIPA), the fee-for-service medical services file and the pharmaceutical services file. These five databases are linked at the source using the health insurance number (HIN), which is then encrypted so that the INSPQ receives a unique number for each person that is different from the HIN.

[1] Philippe Gamache, Institut national de santé publique du Québec, 945 Avenue Wolfe, Québec, G1V 5B3 philippe.gamache@inspq.qc.ca.

Needless to say, these five administrative files were not selected randomly and the stakeholders involved knew their potential to improve chronic disease surveillance. The hospitalization file alone can be used to identify the most serious diseases, those severe enough to require at least one hospital stay. Combining this file with the medical services and pharmaceutical services files made it possible to also identify chronic diseases that require health services other than hospital services.

The fact remains that the primary use of the five administrative databases was definitely not to monitor chronic diseases. Furthermore, the bases do not have the same administrator so that data management and validation is far from uniform. This context presents an initial challenge for proper chronic disease surveillance. The death file is administered by the MSSS jointly with the Institut de la statistique du Québec (ISQ). Its main role is to contribute to the demographic monitoring of Quebec's population. This file is validated and revalidated and the data quality is excellent. The hospitalization file is managed by MSSS and its primary purpose is to track hospital morbidity and the use of health resources. Here again, data quality is not an issue and the multiple causes of hospitalization make it possible to identify many diseases. In the past, before linkage, the death and hospitalization files were the core of chronic disease surveillance. They are, therefore, known entities.

The three other new files come from the Régie de l'assurance maladie du Québec (RAMQ). While they are administered by the same organization, they have very different primary purposes. The insured persons file is used simply to track insured persons and thus determine which citizens are insured by the public health plan and at what point in time. The FIPA also contains socioeconomic data on individuals, particularly place of residence, age and sex. The primary purpose of the medical services file is to pay physicians based on the medical services provided. For this reason, the medical service code is the most important variable and, therefore, is validated the most thoroughly. However, the diagnostic code is the most important variable when it comes to identifying most chronic diseases, but its validation is not as thorough. In practice, the diagnostic code is sometimes missing. Finally, the pharmaceutical services file is used primarily to reimburse persons insured under the public drug insurance plan. While certain medications are clearly associated with a chronic disease, the pharmaceutical services file does not directly identify the diseases that insured persons have because that is not its primary role.

Differences between the five files also complicate identification of chronic diseases or the comparability of estimates over time. First, the shift from the 9th version of the International Classification of Diseases to the 10th version was not done simultaneously for all files. In Quebec, the changeover took place in 2000 for the death file, in 2006 for the hospitalization file, and it has still not happened for the medical services file. The diagnostic codes for the latter are still coded according to ICD-9. Determining the impact of this lack of uniformity on estimates is a challenge that had to be considered and examined.

In addition, the death file is slightly later than the other files. As noted above, this file is rigorously validated, which delays dissemination to users. Certain deaths (mostly accidental) require the involvement of the Coroner's Office. Coroner inquests can take several months—or even more than a year—thus delaying the file's release. As a result, any chronic disease surveillance indicator that utilizes the death file covers a shorter and less recent period. The FIPA also contains a date of death, but no cause of death.

The last major difference among the files is the population covered by each of them. While the death and hospital visit files cover the whole population, without exception, the FIPA and the medical services file cover only Quebeckers insured under the health insurance plan. This means that data on certain groups within the population, such as members of the military, persons living outside Quebec for most of the year, and people with access to the private health system, are not available. Finally, at the time of writing, the pharmaceutical services file covered only Quebeckers aged 65 and older insured under the public plan, excluding individuals covered by private insurance as well as those living in institutions who receive their medication directly through the institution. The lack of uniformity in population coverage creates challenges, first and foremost for the production of population-based estimates from databases that are not entirely population-based.

2. Data quality

2.1 Unequal quality of files and of variables

As noted earlier, all variables of the death and hospital visit files are meticulously validated. Quality is much more variable for the three RAMQ files, not only across the files but within each one. There is no consistency in the reliability of the data from one variable to another, which can obviously impact the estimates produced. Let's use the example of the medical services file again, in which the service code is of excellent quality because the payment of physicians largely depends on this variable. In other words, both RAMQ and physicians have an interest in medical services being properly and accurately coded. This is not necessarily the case for the diagnostic code that accompanies the service code because the diagnosis has no direct impact on physician payment (with exceptions). In the example below, it is possible to identify the chronic disease of the first patient but not the second patient given that the diagnostic code is missing.

Table 2.1.1
Example of unequal quality among variables in the same file

Service code	Diagnostic code (ICD-9)	Disease
09162 – Primary visit	250	Diabetes
09162 – Primary visit	?	?

This example shows the direct impact that a single variable can have on the quality of the estimates produced. The next section highlights other situations in which variable quality is problematic but where the impact is sometimes indirect. A number of inconsistencies among the files will also be explained.

2.2 Quality issues: examples and inconsistencies

One of QICDSS's most important variables is the postal code of the individuals in the insured persons file. The postal code is used to identify each person's place of residence and to track their moves over the years. Finer estimates by health region and by territory are highly dependent on the postal code. However, the quality of this variable leaves much to be desired. First, at the beginning of the period (especially in 1996), many postal codes are simply missing. Then, throughout the period, a small but not insignificant proportion of postal codes are incorrect, most of the time due to capture errors. Lastly, upon occasion, certain "gaps" in the address prevent identification of an individual's place of residence for a certain period of time. More specifically, the end date of one address (postal code) does not correspond to the start date of the next address.

There are also problems with the municipal code, also known as the census subdivision code. RAMQ assigns this municipal code based on the postal code. In a rural setting, several municipalities can have the same postal code. In these situations, the assignment process used by RAMQ assigns the postal code of the most populous municipality. As a result, based on the FIPA file received by INSPQ, about 250 municipalities have no residents. To address this problem, INSPQ developed its own method of assigning municipal codes using the name of the municipality which is also a variable in the file. Unfortunately, the name of the municipality is also a variable of questionable quality. RAMQ does not standardize the names so that there are currently 1183 different municipal names for Montréal residents and 674 different municipal names for residents of Quebec City. Variations include simple typing errors (e.g., Monréal), names of municipalities prior to amalgamation (e.g., Dorval), added prefixes or suffixes (e.g., Montréal QC), almost complete addresses (e.g. Sherbrooke Street, Montréal) or combinations of these factors. INSPQ has to apply some standardization, when possible, in order to work with these different names for municipalities.

Some inconsistencies also affect data quality and, by extension, the quality of the estimates produced. For example, the date of death found in the insured persons file is not always the same as that in the death file. When possible, the latter takes precedence given the more extensive validation of the death file. RAMQ does not require a very precise date of death given that the important thing is knowing that an individual has died and therefore is no longer eligible for the public insurance plan. Another example is that pharmaceutical services sometimes appear for individuals who are not eligible based on the eligibility variable of the public insurance plan. This type of inconsistency is usually explained by administrative delays. Thus, the eligibility variable is not entirely reliable. Third, there is some underestimation of the 20- to 30-year-old population in the FIPA for the simple reason that young adults renew their health cards less often. They no longer have their parents to do it for them and they are generally in good health and not using health care regularly. Lastly, the annual update of QICDSS data in the spring of each year can slightly alter the estimates produced using the system's previous versions, even those from quite far in the past. For various reasons, RAMQ applies these changes to the data (additions, withdrawals, modifications or corrections). It should be noted that the death and hospital visit files do not change–only one year of data is added.

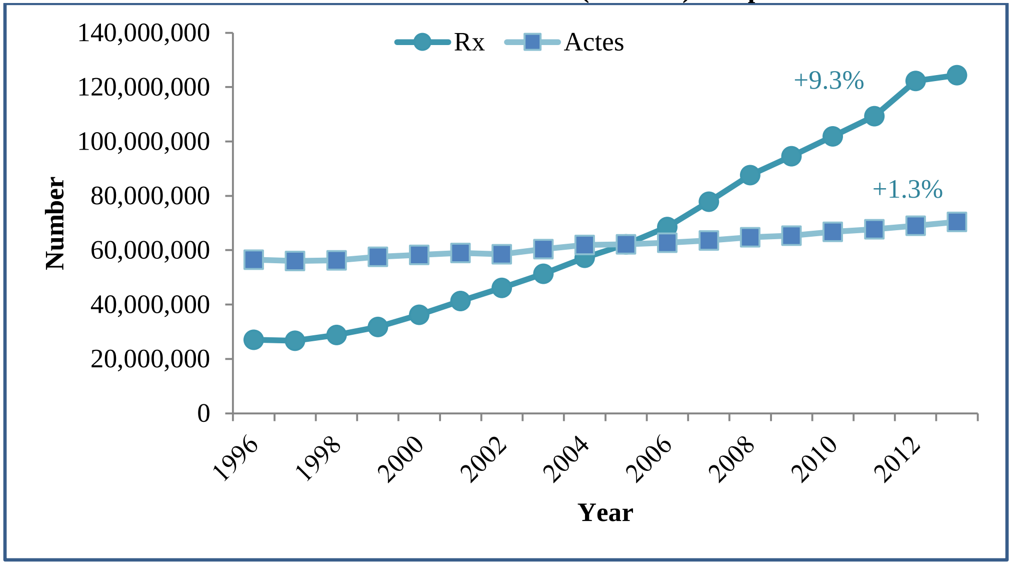
3. Simultaneous use of data

3.1 Size

To some degree, administrative data represent the first big data. They are not the result of a pre-planned survey by one or more researchers but rather the outcome of daily human activity. In the case of interest to us, each health event (death, hospital visit, consultation, drug purchase) is recorded for the vast majority of Quebec's 8 million residents and has been for almost 20 years (1996 to 2014). The size of the data only increases with time leading to certain challenges with their use and analysis.

Over this period, there have been close to 900,000 deaths, 19.5 million hospitalizations, 1.13 billion medical services and 1.22 billion pharmaceutical services, each representing a detailed observation of several variables. These data are stored on a local server and accessed using the SAS Enterprise Guide software. For deaths and hospitalization, the read time for the complete databases, without making changes, takes only a short 6 to 20 seconds respectively. However, the wait time for the complete medical and pharmaceutical services files is currently 75 and 120 minutes respectively at this time. Subsequent manipulations of these files also require a great deal of time. The chart below shows that this situation will continue despite technological and computer advances.

Table 3.1.1
Annual increase in the number of medical services (Services) and pharmaceutical services (Rx)



The wait time is not the only consequence of the quantity of information to be processed. The data also need to be stored, and storage space must be managed appropriately in a context of a relatively large number of users.

3.2 Strategies to avoid lost time and storage problems

If a user is to work efficiently, waiting hours each day simply to process original files is not an option. Thus, it is crucial to create permanent, transformed databases. These databases retain only the observations needed to carry out one or more projects. For bigger tasks, using dead times, such as noon hour, evenings and weekends, is the key to being more efficient. This is especially true given that multiple users are connected to the local server on any given day. Using dead times also avoids slowing down other users by running a big request.

However, the creation of permanent transformed data, that are often transitory, can have the undesirable effect of quickly reducing available storage space. It then becomes important to optimize both permanent and temporary (which is not saved when the program is closed) storage space. In an ideal world, the scalability and elasticity of cloud computing would effectively resolve this issue, but Section 4 of this paper will explain why QICDSS cannot use this technology at this time. The best alternative is to increase the size of the storage space on the local server but this solution is not always possible, especially in a context of budget constraints.

In the meantime, a few other solutions to save time and improve efficiency for database users are possible. One of these solutions is to provide one or more additional computers dedicated to big requests or extensive analyses. In this way, while a big request is running on another computer, the user can continue to work at his own station on another project or another component of the same project. As well, for the purpose of preliminary results, working on a sample of the population would speed the process. In general, results obtained from such a sample remain representative given the size of QICDSS. For example, selecting 10% of individuals in QICDSS still involves working with close to one million individuals, a cohort sufficiently large enough for the majority of analyses. Working in this manner avoids having to run preparatory statistical models, such as survival models and multilevel or hierarchical analyses, which involve hours of waiting.

Lastly, there are two more technical elements that can avoid many hassles. First, proper planning of maintenance time with the information technology team is critical to avoid unpleasant surprises. Maintenance periods at regular intervals make planning much easier for data users. As well, the reliability of the power supply must be excellent to avoid shutdowns at the wrong time. This situation has occurred a few times at INSPQ as its Quebec City offices are unfortunately located in an area where the power supply is particularly sensitive to weather conditions.

4. Other constraints

4.1 Legal constraints

QICDSS was created as a result of a tripartite agreement between INSPQ, MSSS and RAMQ, and was approved by the Commission d’accès à l’information. Any amendments to the agreement, even minor ones, must be approved by all of the organizations. This means that adding information—whether a single variable like the accident code to identify a trauma or the inclusion of all pharmaceutical services, even for those under 65—is a long, difficult process that is often subject to delays. Furthermore, QICDSS can only be used for the purpose of monitoring chronic diseases. Thus, it cannot be used to study infectious diseases or traumas, to give just two examples.

The agreement details all of the data processing steps, primarily to ensure the security of the data and the confidentiality of Quebeckers. Thus, the transfer of data from RAMQ to INSPQ is governed by certain rules as are the warehousing and storage of those data. This is why it is currently impossible to consider cloud computing as a storage solution. At present, the agreement stipulates that the data must be stored on a secure local server to maximize the security of the information. For the same reason, physical access to the facilities is restricted to employees responsible for monitoring chronic diseases and access to the computer stations is governed by the different user statuses. For example, only a few users have access to sensitive variables such as postal code, date of birth and date of death. Computer access is recorded in the event a problem should arise. Finally, the dissemination of results is also governed by a number of security rules, once again to minimize the risk of confidentiality breaches.

4.2 Other data sources

QICDSS allows the addition of important contextual and territorial variables, most often through the postal code. This is how territories, such as health regions, are assigned to all individuals with a valid postal code in order to produce regional or local estimates. One of the weaknesses of administrative databases is the absence of socioeconomic data. Again using the postal code, it is possible to add this type of information using ecological proxies and the deprivation index. However, adding information about individuals is not allowed. This is why linking QICDSS to survey data is prohibited at this time. This type of linkage would make it possible to take lifestyles such as smoking and nutrition into account in our analyses since lifestyles are closely associated with a number of chronic diseases. Not being able to take them into account is a major challenge.

In Canada, certain provinces are already able to make use of electronic medical files as an additional data source. In Quebec, this is not yet the case. There is no doubt that electronic medical files, if they become accessible in the future, will further enhance chronic disease surveillance. QICDSS, as it is currently configured, is definitely an excellent data source enabling calculation of better estimates than in the past. It is also a relatively low cost source given that the data already exist. In short, the challenges described in this paper are significant but there are so many benefits that no one wants to go back in time.

References

Blais C, Jean S, Sirois C, Rochette L, Plante C, Larocque I, Doucet M, Ruel G, Simard M, Gamache P, Hamel D, St-Laurent D, Émond V (2014), « Le système intégré de surveillance des maladies chroniques du Québec (SISMACQ), une approche novatrice (Quebec Integrated Chronic disease surveillance system (QICDSS), an Innovative approach)», *Maladies chroniques et blessures au Canada*, Vol. 34 no 4, p. 247-256.

Geran L, Tully P, Wood P, Thomas B. (2005) Comparability of ICD-10 and ICD-9 for Mortality Statistics in Canada, *Statistics Canada*.

Pampalon R, Hamel D, Gamache P, Raymond G. (2009) A deprivation index for health planning in Canada. *Chronic Dis Can*; 29(4):178-191.

Régie de l'assurance maladie. (2016) Manuel de facturation – Rémunération à l'acte, Mise à jour 89, Janvier 2015, 874 pages.

Rochette L, Émond V. (2004) Chronic-disease surveillance in Quebec using administrative file linkage. *Proceedings of Statistics Canada's 2014 Symposium*.