Quality Assessment and Validation of Administrative Data Sources in Health Statistics

Tina Zupanič[1], Metka Zaletel[2]

[1] *National Institute of Public Health, Ljubljana, Slovenia; tina.zupanic@nijz.si*
[2] *National Institute of Public Health, Ljubljana, Slovenia; metka.zaletel@nijz.si*

Abstract
Sufficient data quality is essential for production of national statistics, both for survey and administrative data used for statistical purposes. The assessment and validation of the latter can be achieved by different means; one method is by using external sources, which are also administrative sources with different initial purpose of data collection, and are judged to be of better quality. This is often the case for health care statistics, where administrative sources are widely used. The initial administrative database in this case is often composed of information gathered from different sources, e.g. medical doctors, patients (their educational attainment and information on their health history), coroners at the hospitals, nurses, etc. The starting point in our case study was the hospitalisation database, which is an important data source for estimates on health care utilisation and provides estimates on the incidence of some diseases. The analysis and comparison with additional external data sources have been performed. These additional data sources could be collected or compiled for statistical purposes (e.g. population and census data gathered by NSI), external administrative databases (e.g. police transport accidents database) and internal databases (e.g. Causes of Death). Variables of interest are in most cases socio-demographic data (e.g. educational attainment level), which are essential for the analysis, but when of poor quality they can lead to false estimations and, consequently, to wrong inferences. In the paper, steps necessary to evaluate and compare different data sources are presented, and the quality analysis of key variables, essential for national and European statistics, are shown. Based on a detailed quality analysis, external data sources were evaluated and assessed as appropriate or inappropriate data sources.

Keywords: quality assessment, administrative data, data linkage

1. Introduction

Nowadays, it is widely known and accepted that administrative data sources have become the most important data source in official statistics, at least within the European Union. There are certainly several fields of statistics where the application of administrative data sources is not so straightforward. One of these is health care statistics, which is, in most of the member

states, prepared by institutions other than NSIs. At the same time, at the Stakeholders' Meeting on Health Statistics, which was held in July 2015 (Brussels), it was recognized that health statistics has received a high priority on the European agenda, and also its complexity with a multitude of stakeholders and data sources, including big data. In the summary of the meeting, it was stated: "The time for developing further surveys is most likely elapsed and administrative sources are to be used more intensively. In view of this cooperation is needed in order to give a better access to administrative sources to statistical authorities as stated in the revision of Regulation 223/2009.".

A statistician with the insight in the process of preparation of health care statistics can find a little paradox in the process: on one hand, a kind of administrative data sources have been used in the process for decades. Here, different data coming directly from health care providers (hospitals, local health centres, etc.) are taken into account. These data are collected partially directly from patients and partially as the result of the health care process, collected by medical doctors and nurses. On the other hand, no additional administrative sources have been used in the process with the role of complementary data source or the role of improving the quality of the final estimations.

We certainly need to acknowledge that in most member states health care statistics is provided by National Institutes of Public Health or some similar institutions, which are sometimes not very obviously a part of the system of national statistics. Hence, it is hard to apply the legal framework of official statistics and gather administrative data sources to raise the quality level of the original data.

1.1 The Case of Slovenia

The National Institute of Public Health (NIJZ) is, traditionally, a part of the system of national statistics in Slovenia. Therefore, statistical legislation is applied in all processes of the preparation of health statistics. In most cases, the initial administrative database (starting point; hereinafter "initial database") is often composed of information, gathered from different sources, e.g. medical doctors, patients (their educational attainment and information on their health history), coroners at the hospitals, nurses, etc. These data are collected on the basis of

national health care legislative framework. Next, these data enter into the system of national statistics and are merged by additional administrative or statistical data sources (hereinafter "additional data source") with three main purposes, so as to:

(1) enrich the original database with additional information / variables,

(2) raise the quality of the final estimates,

(3) decrease the respondents' burden (e.g., it is not necessary to collect data on patients' educational attainment at the hospital).

1.2 Quality Drawbacks of Data Sources for Health Statistics

It is widely known that there are many quality drawbacks of administrative data sources. Wallgren and Wallgren (2007) show that "fundamental idea behind the register system is that many sources should be used so that a high standard of quality and consistency can be achieved". This fact has been neglected in the preparation of the majority of health statistics in the previous decade, also due to an insufficient legislative framework to overcome this problem.

There exist quality drawbacks of the initial database, mainly connected to the lack of knowledge or time, or stress during the process of data collection. A part of the data is gathered from patients admitted to the hospital. These persons are usually in very stressful situations and do not understand the importance of providing some data, e.g. educational attainment. On the other hand, a part of data is provided by medical doctors and nurses. As explained below with the variable "the cause of death", the coroners in the present system of coroners' service in hospitals do not have enough knowledge to provide the underlying cause of death according to WHO definitions, but usually provide the direct cause of death.

As one might expect, there are also quality drawbacks of the additional data sources. The most important cause is that the additional data sources are collected for different purposes with different methodological definitions and different time spans. A typical case can be shown by the Police transport accident database. The unit in this database represents a traffic accident;

the key variables are the date and circumstances of the accident, the number of injured persons and the seriousness of the injuries. It is necessary to point out that (1) the data are gathered on the spot, (2) and the rescuers are sometimes not qualified enough to judge the seriousness of the injuries. Nevertheless, merging the traffic accident database and the hospitalization database might provide estimates on the undercoverage and overcoverage, and also improve some of the variables.

2. The Integration Phase - Steps to Evaluate and Compare

As described in Wallgren and Wallgren (2007), the integration phase appears where different sources are integrated into a new statistical register. Before that, it is necessary to perform the quality assurance of all data sources that are entering the final statistical register. Here, we should distinguish between the additional *statistical* data source, which has already been edited and approved for the use in the national statistical system (e.g., socio-demographic data), and the additional *administrative* data source. The first is ready-to-use and integrated immediately to fulfil the purposes described above. The second should be examined carefully, including the possibility that the level of quality of data will not be sufficient to integrate them into the final register. The most important quality components that are stressed at that point are: consistency and coherence, over- and under-coverage, timeliness and accuracy.

3. Results

Due to poor quality of some socio-demographic data in the National Hospital Health Care Statistics Database (hospitalization database) and the Perinatal Information System (perinatal database), the population and census socio-demographic data gathered by the SORS were linked to the National Hospital Health Care Statistics Database (hospitalization database) and the Perinatal Information System (perinatal database) on the basis of a personal identification number. After merging, the completeness of variables from the initial database and additional data source was observed.

When observing the marital status, we can observe a substantial decrease in the proportion of unknown/ missing data when adding legal marital status from an additional data source.

Tab. 1: De facto marital status in the perinatal database, original data (NIJZ, 2013 and 2014)

de facto marital status	**2013**	**2014**
single	13.2	11.4
married	30.3	19.7
divorced	0.2	0.2
widowed	0.1	0.0
living in consensual unions	37.7	29.0
unknown/missing	**18.6**	**39.7**
Total	100	100

Tab. 2: Legal marital status in the perinatal database, official data (SORS, 2013 and 2014)

legal marital status	2013	2014
single	54.7	55.6
married	43.1	42.1
widowed	0.1	0.1
divorced	1.9	2.1
same-sex registered partnership	0.0	0.0
unknown/missing	**0.2**	**0.1**
Total	100	100

When observing the highest level of education completed, the unknown/missing data in hospital data is even higher than in the case of marital status. After adding data on the highest level of education completed from official statistics population data, the proportion of the missing data decreases a lot, but is still observed due to the hospitalized population still in the process of education (children, students).

Tab. 3: The highest level of education completed, a comparison of the original (SBO) and official data (NIJZ, SORS 2013 and 2014)

	2013		**2014**	
	SBO	**SORS**	**SBO**	**SORS**
valid value	23.5	80.9	28.5	80.8
unknown/missing	76.5	19.1	71.5	19.2

When compared with the hospital data, the initial perinatal data on the highest level of education completed are more complete. But after adding the official socio-economic data, the total picture is almost fully covered with valid values.

Tab. 4: The highest level of education completed, a comparison of the original (PIS) and official data (NIJZ, SORS 2013 and 2014)

	2013		2014	
	PIS	**SORS**	**PIS**	**SORS**
valid value	67.1	99.0	49.7	98.9
unknown/missing	32.9	1.0	50.3	1.1

A similar analysis was done for occupation in employment.

Another assessment was done on the causes of death (CoD) data. Due to changed methodology in capturing CoD in hospitals from 2013 onward, underlying CoD for those individuals that died during hospitalization was linked from annual CoD database, based on personal identification numbers. From 2013 onward, a direct CoD was captured on the field and not an underlying CoD. Table 6 presents underlying and direct CoD for hospitalized individuals. It can be observed that from 2003 to 2012 hospitals actually captured direct CoD and not underlying CoD, which should be captured according to the methodology in force. The shift in the distribution of underlying CoD is seen between neoplasms and diseases of respiratory system, due to non-compliance of the WHO regarding underlying CoD definition in the years from 2003 to 2012 in the initial database. Based on a detailed quality analysis, the annual CoD data were assessed as a more appropriate source for underlying CoD for those individuals that died during hospitalization, than the initial source.

Tab. 5: Underlying/ direct CoD for hospitalized individuals – captured vs. official data (NIJZ, 2003 – 2014)

	Year of data capture													
	Underlying/ Direct CoD												Underlying CoD	
Chapter of Direct/ Underlying CoD	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013***	**2014***	**2013**	**2014**
1 Certain infectious and parasitic diseases	4.28	4.16	4.76	5.88	6.8	7.41	6.96	7.53	9.37	9.08	8.57	6.96	0.91	1.31
2 Neoplasms	24.32	24.02	22.82	20.47	20.21	20.27	20.48	21.65	21.2	18.42	17.52	19.57	37.8	36.85
3 Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	0.16	0.22	0.17	0.12	0.23	0.16	0.18	0.2	0.31	0.28	0.23	0.2	0.17	0.15
4 Endocrine, nutritional and metabolic diseases	0.33	0.34	0.41	0.25	0.27	0.39	0.32	0.42	0.54	0.63	0.44	0.62	1.99	2.04
5 Mental and behavioural disorders	0.06	0.06	0.07	0.07	0.14	0.07	0.1	0.11	0.08	0.09	0.19	0.11	0.79	0.74
6 Diseases of the nervous system	0.48	0.58	0.57	0.85	0.79	0.84	0.98	1.03	0.9	1.08	0.86	0.97	1.45	1.62
7 Diseases of the eye and adnexa	0.01	0	0	0.01	0	0	0	0	0.01	0	0	0	0	0
8 Diseases of the ear and mastoid process	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0
9 Diseases of the circulatory system	41.84	42.12	41.48	40.38	36.76	36.18	34.96	35.25	34.85	35.66	37.86	38.61	33.65	36.15
10 Diseases of the respiratory system	14.13	13.94	14.51	13.99	15.45	15.92	18.7	18.64	19.21	20.9	19.37	18.93	9.04	8.29
11 Diseases of the digestive system	5.17	4.64	3.31	4.02	4.29	4.19	3.48	4.25	5.07	4.84	3.24	3.12	8.8	8.1
12 Diseases of the skin and subcutaneous tissue	0.03	0.03	0.03	0.02	0.02	0.06	0.05	0.04	0.11	0.12	0.12	0.05	0.25	0.08
13 Diseases of the musculoskeletal system and connective tissue	0.16	0.07	0.04	0.05	0.13	0.11	0.09	0.11	0.16	0.15	0.11	0.03	0.49	0.63
14 Diseases of the genitourinary system	0.68	0.68	0.78	0.87	1.26	1	1.08	1.09	1.61	1.46	1.19	1.2	2.32	1.94
15 Pregnancy, childbirth and the puerperium	0	0	0	0	0	0	0	0	0	0	0	0.01	0	0.01
16 Certain conditions originating in the perinatal period	0.28	0.25	0.11	0.11	0.09	0.1	0.03	0.01	0	0.02	0.03	0.01	0.01	0.01
17 Congenital malformations, deformations and chromosomal abnormalities	0.16	0.09	0.08	0.09	0.08	0.05	0.02	0	0.01	0.01	0.03	0.03	0.21	0.26
18 Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	7.81	8.7	10.7	12.66	13.31	13.04	12.45	9.54	6.42	7.06	9.85	9.27	0	0.01
19 Injury, poisoning and certain other consequences of external causes	0.1	0.09	0.14	0.15	0.16	0.18	0.08	0.12	0.14	0.13	0.15	0.18	0	0
20 External causes of morbidity and mortality	0	0	0	0	0.01	0.01	0.02	0.01	0	0	0	0	2.09	1.82
21 Factors influencing health status and contact with health services	0	0	0.01	0	0	0	0	0	0	0.04	0.23	0.09	0	0
22 Codes for special purposes	0	0	0	0	0	0	0.01	0	0	0.03	0	0.01	0	0
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100

The fourth case presented is quality assessment of merged output, where police transport database was linked to in-patient data. In this case, police transport data was the main database (it includes all traffic accidents). Used in-patient data include only more severe injuries (those that needed to be hospitalized) and not those which were only treated in EDs'. Police data were used for the identification of all cases of traffic accidents and, after that, in-patient data were linked to the police transport data. Individuals treated in ED (not captured on individual level) and those that died later than 30 days after the accident occurred (not included due to methodology of case definition) were not linked to the police data. Before linking both databases, transformation of ICD-10[1] codes of injuries and poisonings to Abbreviated Injury Scale (AIS) code was done (based on mapping table received from the European Commission). Because ICD-10 codes included in the in-patient database are only 4 digits long and not as detailed as ICD codes in the mapping table, less codes were transformed to severe injuries AIS code than it was expected (more severe injuries are encoded on lower code position). After that in-patient data were linked to the police data according to the three common variables: date of birth, date of accident/date of injury and gender (no personal identification number available in police transport database). The table below shows percentages of the outcome of transport accidents in police transport database.

Tab. 6: The outcome of transport accidents in police transport database (Police, 2013)

		Percent
Accident outcome	missing	0.4
	material damage	61.4
	with severe body injury	1.7
	with less severe body injury	19.6
	fatal accident	0.3
	without injury	16.6
	Total	100.0

[1] International Statistical Classification of Diseases and Related Health Problems

After linking both databases, results were disappointing. Only 4.9% of all transport accidents actually linked with in-patient data. When observing only accidents with injuries, 22.4 % linked with in-patient data. Linked cases showed different injury severity in both databases (Tab. 7). Results can be attributed to the inadequate mapping table and undercoverage – the exclusion of ED treated transport injuries.

Tab. 7: The outcome of linking police transport database and in-patient database, row percentages (Police and NIJZ, 2013)

		AIS - 3 categories			Total
		AIS 1, 2 (less severe injury)	AIS 3, 4, 5, 6 (severe injury)	cannot determine	
Accident outcome (police data)	missing	100.0	0.0	0.0	100
	material damage	86.2	6.9	6.9	100
	with severe body injury	76.2	22.3	1.4	100
	with less severe body injury	98.0	1.4	0.6	100
	fatal accident	33.3	57.6	9.1	100
	without injury	88.2	11.8	0.0	100
Total		90.0	8.7	1.3	100

Hospitalization data were assessed as an inadequate source for the severity of injury, due to insufficiently detailed information at the level of diagnosis in hospitalization data, undercoverage (not all traffic injuries are included) and also a different purpose of in-patient database.

4. Conclusion

At the beginning, the paradox in the process of preparation of health care statistics was mentioned – administrative data sources have been used in the process for decades, but almost none of them have been used as additional sources to improve the final estimates. Now, we can conclude that even in the case of an additional statistical source with approved quality level, caution is needed due to different usages of the data. On the other hand, sometimes

promising additional administrative data sources turn out to be useless due to a completely different purpose.

Finally, it is necessary to point out that health care statistics could raise the quality level by using more additional administrative and statistical data sources, but joint action at the level of EU would be welcome to spread the knowledge and experiences among countries. Further analyses would enlighten pros and cons in health care statistics.

5. References

European Commission, EUROSTAT (2015). Minutes of the Stakeholders' Meeting on Health Statistics, 8 July 2015. Not publicly available.

Nacionalni inštitut za javno zdravje (2013). Spremljanje bolnišničnih obravnav (SBO). Definicije in metodološka navodila za sprejem podatkov o bolnišničnih obravnavah preko aplikacije ePrenosi, v 1.3. Available at: http://www.nijz.si/sites/www.nijz.si/files/uploaded/mg_sbo_mn_eprenosi_v1_3.pdf

Nacionalni inštitut za javno zdravje (2013). Perinatalni informacijski sistem RS (PIS). Definicije in metodološka navodila za sprejem podatkov perinatalnega informacijskega sistema preko aplikacije ePrenosi, v 1.5. Available at: http://www.nijz.si/sites/www.nijz.si/files/uploaded/mg_pis_mn_ver_1_5.pdf

Wallgren, A. and Wallgren, B. (2007), Register-based Statistics. Administrative Data for Statistical Purposes, J.Wiley, N.Y.