Charles University in Prague

Faculty of Social Sciences
Institute of Economic Studies

MASTER THESIS

Quality and performance assessment of healthcare providers in Slovakia on the basis of administrative data

Author: **Bc. Tamara Vraždová**

Supervisor: **Mgr. Henrieta Tulejová**

Academic Year: **2012/2013**

Declaration of Authorship

The author hereby declares that he compiled this thesis independently, using only the listed resources and literature.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, January 7, 2012

Signature

Acknowledgments

The author is grateful to Dôvera, zdravotná poisťovňa a.s. for granting access to data essential for data analysis in this thesis. All data were anonymized – at both the patient and hospital level. Providers were assigned a meaningless identification number in the analysis and cannot be identified.

Abstract

The aim of this thesis was to explore options for hospital profiling in the Slovak republic. Sacrificing breadth of the study in favor of depth, the scope of the analysis was narrowed down to one quality indicator only – mortality. In the first step a mortality prediction model was constructed in order to predict expected probability of death on the basis of a set of risk factors in order to filter away variation in hospital outcomes that is caused by other factors than quality of care. Validation of the model was performed on a validation sample of 25% of data. Discriminative ability of the final model is very high – c-statistics over 0.9. Furthermore, we verified that hospitals differ in the risk structure of their patient populations significantly – mean predicted probability of dying for hospitals differed from 0.02% to 33%. In the second step hospital profiling was performed. Standardized mortality ratios were calculated for each hospital as a difference between observed and expected number of deaths. After introduction of risk-adjustment and calculation of confidence intervals 43% of hospitals were re-classified. 30-day mortality was selected as best indicator for hospital profiling.

Keywords

Hospital profiling, risk-adjustment, performance and quality assessment, public performance reports, healthcare, Slovakia

Abstrakt

Cieľom práce je preskúmať možnosti hodnotenia poskytovateľov zdravotnej starostlivosti na Slovensku. Z dôvodu obmedzeného priestoru bol rozsah práce zúžený iba na jeden indikátor kvality – úmrtnosť. V prvom kroku bol zostavený predikčný model úmrtnosti za účelom predikcie očakávanej úmrtnosti na základe niekoľkých rizikových faktorov. Tento model nám umožní odfiltrovať variabilitu v úmrtnosti, ktorá je spôsobený inými faktormi ako kvalitou poskytnutej zdravotnej starostlivosti. Validácia modelu prebehla na 25% validačnej vzorke dát. Diskriminačná sila zostaveného modelu je veľmi vysoká – c-štatistika dosahuje hodnoty nad 0.9. Zároveň sme overili, že existujú výrazné rozdiely v rizikovej štruktúre pacientov jednotlivých nemocníc – priemerná predikovaná pravdepodobnosť úmrtia sa medzi nemocnicami výrazne líšila v rozmedzí 0.02% až 33%. V ďalšom kroku sme pristúpili ku klasifikácii nemocníc. Štandardizovaná miera úmrtnosti bola vypočítaná pre každú nemocnicu ako podiel skutočného a očakávaného počtu úmrtí. Po zohľadnení rizikovej štruktúry pacientov a výpočte konfidenčných intervalov bolo preklasifikovaných 43% nemocníc. 30-dňová úmrnosť bola vybraná ako najvhodnejší indikátor úmrtnosti pre účely hodnotenia poskytovateľov.

Klíčová slova

Hodnotenie nemocníc, riziková štruktúra pacientov, hodnotenie kvality, verejné hodnotenia, zdravotníctvo, Slovensko

Bibliografická evidence

VRAŽDOVÁ, T. (2013) *Quality and Performance Assessment of Healthcare Providers in Slovakia on the Basis of Administrative data*. Praha, 2013. 126 s. Diplomová práce (Mgr.) Univerzita Karlova, Fakulta sociálních věd, Institut ekonomických studií. Vedoucí diplomové práce Mgr. Henrieta Tulejová

Contents

List of Figures

Master Thesis Proposal

Author:	Bc. Tamara Vraždová
Supervisor:	Mgr. Henrieta Tulejová
Proposed Topic:	Quality and performance assessment of healthcare providers in Slovakia on the basis of administrative data

Topic characteristic

In my thesis I will address the issue of quality measurement of healthcare providers. The aim is to examine approaches of how to identify differences in the performance of individual providers.

When comparing performance of hospitals we encounter a major problem of separating various factors that affect the results they report: structure of the patients treated by the establishment, pure chance (bad luck – sometimes the patient dies no mater what you do) and finally the actual treatment provided by the healthcare provider. In order to be able to differentiate between healthcare providers and set the purchasing strategy accordingly – rewarding those providers that offer higher quality and better value health care, we need to filter away the first two influences. In particular, we need to employ risk-adjustment on the data reported to the insurance company in order to control for the particular mix of patients treated by the healthcare provider. Typically, we standardize the data for age, sex and risk factors that describe health status of the patient – diagnosis, comorbidities and/or complications at the admission. Furthermore, after dealing with the issue of risk-adjustment, it is necessary to employ statistical techniques to treat the problem of pure chance and small numbers often encountered in the health data. This is often done by calculating confidence intervals, in this particular problem Bayesian interval estimates are most commonly used.

Issues described above are well treated in Anglo-Saxon literature. Further problems, however, arise when we attempt to employ this analytical framework in Slovak hospitals. Due to the specific nature of hospitals' reporting process to the health

insurance companies, there is virtualy no information on particular procedures provided to patient, drugs administered during his stay in the hospital or the actual times between admission and treatment, etc. In other words, once the patient is admitted to the hospital, the purchaser looses track of any and all treatment provided to the patient other than getting rough information about his movement between hospital departments. Even more troublesome is the quality (or lack thereof) of those few information that are being reported to the insurance companies – namely diagnosis and comorbidities, both of which are fundamental for risk-adjustment. It will be the aim of my thesis to search for ways of how to deal with these problems and gain the best possible estimate of providers‘ quality given the data available at the moment and how it could evolve in light of envisioned introduction of DRG payment mechanism.

Hypothesis

1. Risk-adjustment for age and sex at the hospital-wide level should improve accuracy of the providers' evaluation and comparison
2. Risk-adjustment improved by diagnosis and some proxy of comorbidity should further improve accuracy of the assessment

Methodology

In my thesis I will endeavour to examine the options for assessment of healthcare providers in the specific setup of Slovak healthcare. Using administrative microdata of insurance company, I will examine various models of risk-adjustment and compare their performance. These models will differ by the selection of variables – the simplest model will include only age and sex and then I will search for the best proxy of capturing the patient's health status such as major diagnosis, mode of hospital admission – acute or planned, transfer of patient to or from other hospital (since transfer can be expected to be related to the severity of the patient's condition), length of stay and previous treatment that might indicate comorbidities. Relevant conditions will be identified on the basis of previous hospitalizations and/or outpatient medications, such as PCG classification. In order to identify relevant conditions, we will use Charlson comorbidity index. Risk-adjustment will be performed by comparing expected and observed values of the variable of interest, where expected values are calculated via hierarchical regression model for the measures of mortality and rehospitalization and linear regression in case of length of stay and additional

costs. Comparing expected and observed numbers we can make assesment of whether the hospital performed better or worse than others.

Additionally, due to problems caused by unreliable and insufficient coding of data I will search for specific hospital departments or diagnosis which are generally characterised by limited variability of treated cases – thus the risk-adjustment we apply should be sufficient to filter away unwanted influences and let us get a more accurate picture of the quality differences between healthcare providers. Indeed, it is a common practice to use diagnosis or procedure specific indicators, such as mortality on acute myocardial infarction. We might, however, not be able to go into such a level of detail because of the problem of small numbers.

Outline

1. Assessment of healthcare providers – literature review
2. Assessment of healthcare providers in Slovakia – current practice and problems
3. Data analysis
    i. Development of predictive model of hospital mortality
    ii. Calculation of the indicators with and without risk-adjustment for individual hospitals (based on the comparison of observed and expected mortality, the latter calculated based on the model developed in part a))
    iii. Searching for patterns in the data that would indicate areas where the performance is primarily influenced by the quality of healthcare provided
    iv. Dealing with problem of small numbers – calculating confidence intervals and possibly some other adjustments
    v. Calculation of quality scores for healthcare providers – summary measures

Core bibliography:

1. Canadian Institute for Health Information (1991). „Technical Notes. Hospital Standardized Mortality Ratios.“
2. Cerrito, Patricia (2010). *Text Mining Techniques for Healthcare Provider Quality Determination. Methods for Rank Comparisons.*New York: Medical Information Science Reference.
3. Ministerstvo zdravotnictví ČR (2005). *Návrh a zhodnocení ukazatelů kvality a výkonnosti akutní nemocniční péče založených na administrativních datech.* Available online at <>

4. Dr. Foster Intelligence (2012). "Understanding HSMRs. A Toolkit on Hospital Standardized Mortality Ratios.Version 7.*"* Available online at <>
5. Centers for Medicare and Medicaid Services (2011). „Medicare Hospital Quality Chartbook. Performance Report on Readmission Measures for Acute Myocardial Infarction, Heart Failure and Pneumonia." Available online at <>
6. JCAHO. Developing Risk Adjustment Techniques Using the SAS Systém for Assessing Health Care Quality in the IM System. Available online at http://www2.sas.com/proceedings/sugi22/POSTERS/PAPER225.PDF
7. Shahian, David M. & Sharon-Lise T. Normand (2008). "Comparison of "Risk-Adjusted" Hospital Outcomes". *Circulation*. 2008; 117: 1955-1963.

Author

Supervisor

Introduction

Profiling of hospital providers and making the results publicly available is now considered a standard practice in much of the Western world. This trend is result of a growing sense of urgency regarding the accumulation of evidence that quality of care provided differs considerably between healthcare providers and yet, there is no way of telling apart good and bad provider other than informal word of mouth. While there are many explanations of why these deficiencies in delivery of healthcare exist, „*lack of a transparent, explicit, systematic, data-driven performance measurement and feedback mechanism for healthcare providers has been considered to be a major contributor*" [Lansky, 2002]. Many arguments are put forward of how performance assessment might help rectify this situation: empowerment of consumers in choosing the best doctor/hospital that would in turn put market pressure on low-quality providers, provision of crucial information for third-party payers and governmental officials to enable them to identify under-performing areas/providers and act accordingly or release of a benchmark and feedback mechanism for the providers themselves – you can hardly expect providers to invest time and money into quality improvement if they are unaware of the lack of quality in the first place. Regardless of the fact that there are many caveats to these arguments and the impact of public performance measurement initiatives is often less than straightforward, the fact remains that some measure of transparency and accountability has to be introduced into previously closed-off healthcare sector.

The task of compiling of hospital profiling is by no means an easy one. Outcomes of care are influenced by three factors: severity of the condition, quality of treatment and pure chance – random variation. All of these factors need to taken into account and dealt with if we wish to achieve a meaningful comparison of quality between providers that truly reflects differences in quality only. Filtering away unwanted variation in the hospital outcomes requires application of advanced statistical techniques and careful validation of all the methods and results.

The primary aim of this thesis is therefore to explore options for hospital profiling in the Slovak republic. This involves selection and implementation of standard statistical techniques used for quality measurement and adoption of these techniques to the particular setting of Slovak healthcare sector, taking into account limited data availability and local coding practices. Sacrificing breadth of the study in favor of depth, I will focus on a single quality indicator – mortality in hospitals.

First part of the thesis provides an introduction into the issue of public performance reporting and brief description of the current state of affairs in Slovakia and peculiarities of the local health care sector that need to be considered (Chapter 1 and 2). In the second part of the thesis, I will overview major methodological considerations in hospital profiling (Chapter 3) and move on to constructing a hospital ranking using data on all hospitalizations in the period January 2010 – June 2012 as recorded by single private health insurance company[1] that covers roughly 30% of Slovak market. An attempt was made to conduct a full hospital profiling from the moment of handling and cleaning data and running a preliminary data analysis (Chapter 4), selection and validation of the best risk-adjustment model and mortality indicator (Chapter 5) and calculation of confidence intervals in order to finally classify hospitals as high, average or low quality (Chapter 6). However, due to the extent of the task, particular focus was paid to developing a risk-adjustment mechanism that would be able account for the particular structure of patients treated by individual hospital providers. Accommodating the other major source of variation – randomness, was not done in full extent and further research is required.

[1] Author is grateful to Dôvera zdravotná poisťovňa, a.s. for kindly providing access to data necessary

1. Introduction into public reporting of health care quality and performance data

1.1. Why get involved in (public) performance and quality reporting?

Leading U.S. agency on healthcare research AHRQ estimated that 1 in every 10 patients who died within 90 of surgery did so as a result of preventable error, costing the system additional $1.5 billion annually [AHRQ, 2008]. Study released by the Institute of Medicine found that between 44 000 and 98 000 Americans die each year in the U.S. hospitals due to preventable medical errors, making hospital errors between fifth and eight leading cause of death in the U.S. killing more people than breast cancer or traffic accidents [IOM, 2012]. These numbers cannot be taken at the face value, they are very rough estimates that depend heavily on the definition of preventable medical error and ability to identify single one cause of death, nevertheless they draw attention to the fact that quality of the healthcare provided in hospitals is seriously lacking even in arguably the most expensive and advanced health care system in the world. While there are many explanations why this is so, „*lack of a transparent, explicit, systematic, data-driven performance measurement and feedback mechanism for healthcare providers has been considered to be a major contributor*" [Lansky, 2002]. This is true in case of fragmented decentralized marketplace where majority of health care providers are private entities as well more centralized systems where state is still a major player. In this section I will review main arguments in favor of public performance ratings.

1.1.1. Quality improvement

Past two decades have witnessed proliferation of various public releases of comparative data on health care providers' performance and quality, primarily in the English-speaking countries of US, Canada, Australia and UK. These initiatives have been guided by the notion that making information on performance publicly available should put pressure on health care providers to increase quality of health care. There

are two main channels through which this can happen. First of all, disclosure of comparative information is expected to reduce information asymmetry and create pressure from the bottom – informed consumer who will seek to choose the best providers once he/she is provided with reliable information on health care quality. Aggregated decisions of these pro-active consumers who select higher quality providers should cause market share of the under-performers to shrink down accordingly, forcing them either to change and improve or driving them out of business. Second channel through which quality improvement might be encouraged is provision of objective external feedback to the providers themselves that helps them reveal existing problems and identify potential for improvement in underperforming areas. Comparative feedback might also assist management in their efforts to implement internal changes within the organization – if the circumstances are right, these data might become powerful argument for persuading reluctant staff that changes are needed [Marshall et al., 2000].

Both of these pathways operate via the concern of health care providers about their public image and market share. Therefore, public pressure might not work if a competitive market does not exist or if the provider has a dominant position in the region (such as a large hospital with no viable competitors in a reasonably close radius). In these cases, pressure on quality improvement might be exerted by third-party payers who pay the bill in most cases. They might condition reimbursements for health services on certain criteria, in more extreme cases even to employ practice of selective contracting, discontinuing contracts with continually low-performing health care providers. This is the basis of pay-for-performance contracts when the payments for health services are not automatic, but certain standards need to be met by the providers. Of course, in this case performance data need not necessarily be publicly available. However, combination of financial incentive with concerns about reputation damage might be particularly effective in stimulating desired changes. Indeed, practice of implementing pay-for-performance initiatives have increased use of performance results among purchasers (Grumbach et al, 1998; Rosenthal et al, 2005).

However, while public reporting of information might be ineffective when the health care market is not sufficiently competitive, it may also foster more competition. Not only by highlighting differences between individual providers and placing them at the mercy of public opinion and market forces but public rating might also serve those same providers as powerful marketing tools facilitating the process of attracting

consumers/purchasers as well as recruiting and retaining high quality staff [Mannion – Goddard, 2003].

1.1.2. Increasing public accountability and facilitating external regulation of the system

Regardless of potential positive effects on quality improvement and cost containment, public disclosure of hospitals‘ data is crucial for increasing accountability of the entire health care system, which is largely paid from public money in most countries. Especially so in case of large faculty hospitals in Slovakia which are still publicly owned and where any kind of accountability is regrettably lacking. It might also provide crucial data for identification of national health care priorities, guiding national public health policies and highlighting areas in need of regulation. Last but not least, these comparative measures might provide invaluable data for epidemiological and other public health data for clinical research. [Mannion – Goddard, 2003]

1.1.3. Cost-effectiveness improvement

Other than quality improvement, publication of comparative performance results might also serve as a tool for cost control though this is usually not the primary aim.

1.2. Review of empirical evidence

In the previous section we reviewed main theoretical arguments for public disclosure of hospitals‘ performance measurements. But, crucial question is, do these theoretical assumptions hold in real life? Empirical evidence on the actual effect of the public reporting and performance data is mixed at best. There are several aspects that need to be considered:

- Do people/physicians/hospitals/purchasers use the reports? Does reporting change how and which hospitals are selected by the consumers?

- Does the reporting have effect on activity exerted by the hospitals? Change of behavior, processes?

- Most difficult question, does the reporting have effect on the outcomes of care? On the quality?

- How reliable are performance and quality indicators? Do they report consistent results?

1.2.1. Focus on consumers: does public reporting affect selection of hospital?

Public disclosure of hospitals‘ data is aimed primarily at the final consumer of health services – the patient. Evidence up-to-date suggests that despite the fact that consumers do want more information on the providers‘ performance (Edgman-Levitan and Cleary, 1996; Hibbard and Jewett, 1997; Robinson and Mollyann, 1997), it has only very limited impact on their decision making.

In one of the early reports, Mennemeyer et al. (1997) examined the impact of Health Care Financing Administration mortality rates on utilization rates on individual hospitals. He found very small effect – hospitals that had twice as high mortality rates as expected recorded one less discharge per week following public release of the data. Interestingly, he also noted that press publication of anecdotal story of one death in hospital had a rather large effect on the utilization – 9% reduction in hospital discharges in the following year.

A number of papers (Hannat et al, 1994; Chassin, 2002) focused on the New York State Cardiac Surgery Reporting System – one of the first public reports available and one of the best documented one up-to-date. None of the papers found any significant impact on the volume of CABG surgeries performed in the low or high performing hospitals. Furthermore, Schneider and Epstein (1998) went deeper and investigated decision making at the consumer level. They found that as little as 12% of patient undergoing CABG surgery were aware of the existence of performance reports and only 1% admitted that these reports had any influence on their choice of surgeon/hospital.

In the same vein, papers studying other reporting systems (Baker et al, 2003; Hibbard et al., 2005) found no meaningful impact on the rate of utilization. One of the few studies that actually identified some impact - Romano and Zhou (2004) found only

short-term change in the volume of CABG procedures performed in both the low and high mortality outliers. [Shekelle et al, 2008]

A number of reasons have been proposed as possible explanations of these less than encouraging findings. Mannion and Goddard (2003) summarized the proposed alternatives: *„a limited window of opportunity to search for clinical information between the onset of illness and the need for health care (Hibbard et al 1996); comprehension problems, for example interpreting whether high or low rates on an indicator show good performance (Vaiana & McGlynn 2002) and difficulties around understanding technical terms and quantitative data (Robinson & Brodie 1997); a general lack of 'trust' in the data provided by government agencies (Bentley & Nash 1998); a predilection for making decisions on the basis of informal information supplied by family and friends, rather than official documents (Mennemeyer et al . 1997); fatalism, 'when it is your time to die there is nothing anyone can do' (Hibbard & Jewett 1997); and a lack of motivation stemming from a perceived limited choice of alternative providers within a reasonable travelling distance of home (Schneider & Epstein 1998)*".

While all of these might be right to some extent, most of the aforementioned problems are probably caused by irrelevancy of reported indicators to its intended recipient – the consumer. Experience shows that consumers define „quality" in different terms than the experts. For example, Kaiser Family Foundation found that consumer is most of all interested in financial affordability, doctor's professional qualifications and accessibility of care [Hibbard and Sofaer, 2010]. Instead of these information, consumer is presented with information on 30-day in-hospital risk-adjusted mortality on CABG or performance of the hospital on the indicator that monitors administration of beta-blockers within 30 minutes of admission for cardio patients. Additionally, he is cautioned to consider these data cautiously because there is always a degree of uncertainty in statistical data. An effort should be made to bridge this gap between needs of consumers, which are not unreasonable, and the actual content of report cards.

Even more disconcerting is the second major finding of the research on the preferences of consumers – rather than technical information compiled by experts, consumers give preference to ratings by other patients – patient satisfaction surveys. Even these, however, are not ultimately reliable source of information, because consumers would like to know opinion of those who are like them – therefore the single most trusted source of information on quality of health care are family and

friends: „*Word-of-mouth information from families and friends was viewed as more trustworthy than summary satisfaction scores*“ [Edgman-Levitan and Cleary, 1996]. Multiple researchers confirmed this finding, including for example Mennemeyer et al. (1997).

This last finding calls into question, yet again, the assumption of classical economy of homo economicus - rational utility-maximizers seeking to systematically improve their well-being. Proposing programs and setting the course of action based on this erroneous paradigm is not only very useless but also very costly. A point in the case: no evidence has been found up-to-date in support of the assumption that you only need to release data on hospital performance, and consumers will do the rest of the work because it makes sense for them to do so. Effort should be made to either incorporate findings from behavioral economics into the design of performance data for the consumer audience (user friendly presentation of information, simplification, etc) or to focus on other audiences altogether.

1.2.2. Focus on providers: does public reporting encourage change of behaviour?

A number of papers (Chassin, 2002; Longo et al., 1997) suggested that health care providers are responsive to publication of comparative data, especially so when the provider is identified as a poor performer in certain areas. Indeed, following release of public data, some hospitals took measures to improve their results, such as implementation of quality assurance programmes, retraining of staff, closer monitoring of performance of physicians, introduction of car seat programs and nursing education in obstetric hospitals [Shekelle, 2008]. Particularly interesting was a set of studies conducted by Hibbard, Stockard and Tusler (2003, 2005) because they compared impact on hospitals that received only confidential feedback or none whatsoever and on hospitals who had their performance results released publicly. They found significant difference in the intensity of quality enhancing activities in favor of those hospitals that were subject to public scrutiny. It seems that concern about public image is a particularly strong driver of positive change in the hospitals, even without external incentives: „*An intrinsic professional competitiveness on the part of the provider organizations, based on a desire to be seen to be performing well, might therefore be just as important a motivator as economic gain.“* [Marshall et al., 2000]

Yet, not all researchers agree on the positive influence of public reporting on quality improvement activities, for example Luce et al. (1996) found no changes in the affected hospitals, neither did Mannion and Goddard (2003), a first study on the effect of public reporting outside of the US. Not only did they fail to detect positive impact of clinical indicators in Scotland, they also investigated why this is so and found two major obstacles: 1) published data were not viewed favorably in terms of credibility and timeliness and 2) lack of incentives in the UK where the ability of purchasers of healh care and consumers to choose alternative providers is rather limited compared to the highly competitive market in the US. This study therefore suggests that publication of performance reports can induce quality improvements as long as the health care market is reasonably competitive.

A different approach was taken by Haffner et al. (2011) who led structured focus-group interviews with hospital staff in 29 randomly selected hospitals in order to find out their perceptions of the practice of publicly reporting performance data. He found that these data: „*i) led to increased involvement of leadership in **performance** improvement; ii) created a sense of accountability to both internal and external customers; iii) contributed to a heightened awareness of **performance** measure **data** throughout the **hospital**; iv) influenced or re-focused organizational priorities; v) raised concerns about **data** quality and vi) led to questions about consumer understanding of **performance** reports"*. It seems that the practice of public reports at the very least succeeded in motivating hospitals to become more aware of their overall performance and performance in relation to their peers. It helped to focus their care improvement activities at those areas that were found particularly problematic. Concerns about the quality of data and problems associated with utilization of these data by consumers, who might misinterpret these, need to be addressed in order to improve impact of public reporting.

Research on the responsiveness of providers to public reporting, however, revealed also dark side to this practice. Commonly cited objection to publication of comparative data states that good results in ratings might be achieved via 3 strategies: 1) improvements in quality 2) rejection of high-risk patients and 3) gaming strategies, of which most important is manipulation of reported data (phenomenon known as „DRG creep" is one of the examples). Only the first strategy is desirable, unfortunately, it is also the most difficult one to implement – it is much more easier to change reported data than to decrease mortality. Papers by Werner and Asch (2005) and Narins et al. (2005) all found evidence about increased reluctance of

doctors to care for high-risk patients following introduction of New Yor*k Cardiac Surgery Reporting System.* [Shekelle, 2008]. Similarly, regarding the third strategy, when studying the impact of public reports on hospital outcomes – mortality specifically, Baker et al. (2002) found that improvements in in-hospital mortality were offset by increases in mortality of discharged patients. Hospitals simply improved their scores by discharging those patients that were about to die. Another option is to keep alive seriously ill patients with no prospect of actually getting better only to escape the clause of 30-day mortality [Normand and Shahian, 2007].

1.2.3. Focus on quality: does public reporting have effect on the outcomes of care?

The final question is the most important one because we argued at the beginning of this chapter that quality improvement is the single most important reason behind the idea of public disclosure of performance data. Not surprisingly, evidence is particularly contradictory on this issue. One of the problems, according to Shekelle et al. (2008), is attributing causality of recorded quality improvements in mortality rates to public release of performance data because at the same time poor performing providers improved their scores, so did everyone else. When this factor was considered, most of studies detected only very moderate or no positive effect on health outcomes. For example, while Hannan (1994) reported reductions in mortality in New York in one paper, in the following he found that mortality rates for ALL hospitals and surgeons improved, even though admittedly, best improvements were observed in the originally worst providers. Similarly Cutler et al (2004) and Baker et al (2002) found no effect on health outcomes once the trend of universal quality improvement was controlled for. Other than that, it were primarily earlier studies that found evidence of positive impact on health outcomes (Longo et al., 1997 for example) who was one of the few ones who focused on a parameter other than mortality following CABG surgery in New York – he found improvement in high-risk infant transfer (implementation of car seat program and formal transfer arrangements) and very low birth weight infants in obstetrics departments. More recent papers failed to confirm this positive impact with the exception of Hibbard et al (2005) who found that obstetrics departments subject to public reporting were more likely to improve scores received on patient safety measures.

These results do not seem overly supportive of the notion that public reporting encourages improvements in health care. However, while the review of evidence revealed negligible effects on the outcomes of care in terms of such complex measures as mortality rates, there seems to be an indication that improvements were recorded in the process of delivery of health care – see studies by Hibbard et al (2005) and Longo et al (1997). Similar conclusion was also reached by Canadian study done by Jack Tu on the effects of the reporting initiative EFFECT (evaluating cardiac treatment) who concluded that: „*We saw some really important changes at the local hospital level and galvanized more than half of participating hospitals to improve care.“* [Tu et al., 2009]. The study analyzed 86 hospitals that were randomly allocated to either receiving early (before publication of the results) or delayed feedback (after public release of the data). Among the improvements identified by the study was reduced door-to-balloon times, better care of cardiac patients in the operating theatres, greater use of ACE blockers and angiotensin-receptor blockers or higher probability of performing rescue PCI (percutaneous coronary intervention).

1.2.4. Validity and reliability of performance measurement. Do the ratings yield consistent results?

There is one remaining issue that needs to be covered regarding the rationale for engagement in the practice of public performance and quality ratings. So far we covered the evidence of whether consumers and providers use these reports (no and yes) and whether they affect the outcomes of care (mixed results). But before making the final decision on the benefits of this practice, we need to consider **how accurately are we able to identify and quantify quality differences between individual providers**. Do the ratings and indicators they are based on truly reflect actual differences between hospitals? There are numerous issues that need to be considered and I will elaborate on these in more detailed in the next chapter on the methodology, but first, let us look at the studies that examined consistency of various performance ratings compiled with different methodologies.

Literature review reveals that different estimates are largely inconsistent. Study by Shahian et al. (2010) compared results of four different models (compiled by four different entities – independent vendors that offer these models commercially to hospitals for internal quality assessment) designed to predict in-hospital mortality for hospitals in Massachusetts based on data of acute care hospitals for three years.

Researchers then compared outputs of these models, primarily focusing on the identification of the best and worst performers and concluded that all four methods yielded substantially different results. Particularly worrisome is the finding that 12 out of 28 hospitals were categorized as having above-average in-hospital mortality by one method while classified as having below-average mortality by other two methods. The study concluded that these inconsistencies might have been caused by either different statistical methods (weak reliability), particularly different exclusion criteria for patients, diagnosis and hospital types that resulted into substantially different patient populations being evaluated (only 1/5 of discharges that were included in all 4 methods) or by the fundamental error in the supposed link between hospital-wide mortality and quality of care (weak validity) (for more details on both main methodological problems that affect the results and using mortality as quality indicator, see methodology section). Another study, similar in design, Stausberg, Halim and Färber (2011) also found unsettling lack of consistency in identification of top and worst performers by different methodological designs: when either the set of indicators or reference values changed, half of the hospitals in the study sample shifted from inferior to superior or the other way around. However, somehow positive was the finding that when various aspects of quality were taken together (patient safety, mortality, length of stay), results were rather robust. Problems started when only one measure of quality was used to judge the overall quality of care in hospitals.

1.2.5. Summary

Results of the literature review differ substantially according to target audience of the reports – consumers or expert audiences - doctors, hospital management, third-party payers. Empirical evidence is rather unambiguous on the effect, or lack thereof, of public performance results on consumers‘ decision making – there is little to none effect, consumers use ratings only sporadically because they do not understand and/or trust the data. There exist strategies to mitigate impact of adverse factors that prevent higher utilization of performance data by consumers – providing useful consumer data (parking spaces availability, fees, waiting times, etc.) together with performance and quality indicators, simplification of the final consumer reports (aggregation of indicators into a few easily understandable quality dimensions), reporting on those aspects of health care quality that interest consumers (doctors qualification, patient satisfaction scores), user-friendly graphical presentation of data (so that they can be

immediately understood) and ensuring high credibility of the data source (independent sources and ideally public endorsement by some credible trusted authority that helps not only with the credibility problem but also provides wider dissemination of the report among consumers).

From this point onwards, however, I will focus on another target audience – the experts. Evidence shows that hospitals are rather responsive to the data, primarily if they operate in sufficiently competitive market, assessment is mandatory and/or good performance is rewarded (such as in P4P schemes). Credibility of the published data is also very important – ideally an independent trustworthy source should be used. If the data are not trusted by the providers, they might either ignore them (as was the case in Scottish public reporting project) or even worse, they might completely block the effort as was the case in 2005 in New York state, when the attorney general's office asked UnitedHealthcare insurance company to halt its planned introduction of providers' ranking [New York Times, 2007].

Impact of the public report cards on quality improvement is more ambiguous, some studies found no improvement in the outcomes of health care, while others identified some limited impact. In agreement with findings that publication of performance data stimulates more performance-oriented activity in the hospitals, process of the delivery of health care appears to improve in response to the feedback provided by the public data. We can therefore conclude that while there is little evidence of quality improvement at the system-wide level, such as mortality, it provides an excellent tool for those providers that do want to improve and they can do so by focusing on processes in the areas identified as problematic by the performance reports. Moreover, increase in the accountability and transparency of previously closed-off sector alone is well worth the effort.

One area of concern remains and that is the lack of consistency of various performance measurement methodologies. Any comparative data need to be released and/or used for selective contracting/rewarding only very cautiously and the methodology used for calculations needs to be assessed for consistency with other reports if available, if not, sensitivity analysis of the indicators must be performed. Holistic approach is preferred – measuring several dimensions of quality rather than only one indicator. Compilation of ranking is also very tricky – while consumers might welcome such clear-cut and easily understandable presentation of information, we saw that even identification of superior and inferior providers fluctuated widely.

Degree of error in case of ranking from the best to the worst provider is bound to be much higher.

2. Description of the current practice of performance and quality measurement in Slovakia

In addition to problems and issues described in the previous section that concern quality assessment in general, implementation in any given region always has to consider local context. In this section, features and issues relevant to hospital profiling in Slovakia will be reviewed.

Situation in Slovakia is unique by the fact that there exists competition both on the health insurance market – multiple health funds operate in Slovakia, and on the health care providers market – health funds can engage in selective contracting and consumer can freely choose his/her provider and yet there is precious little effort at evaluating and comparing these health care actors. Given the severity of information asymmetry problem in health care setting, this is a serious shortcoming because you cannot achieve truly competitive market with all its benefits unless you come up with a way of separating good insurance funds from the bad ones and excellent hospitals from those that are more likely to harm you than cure you. It is aim of this thesis to explore options for correcting this deficiency.

2.1. Current state of affairs

2.1.1. Official framework

Officially, legislative framework that incorporates quality indicators of healthcare providers does exist – §7 of the act no. 581/2004 on health insurance companies specifically mentions quality indicators as one of the parameters health insurance companies are legally obliged to use during the process of contract negotiation with healthcare providers (ranking of the providers is used for setting reimbursements). Quality indicators can also provide legal grounds for discontinuation of contract with healthcare provider in cases when deviation from quality standards is significant and consistent and is confirmed by audit of the quality of delivered health care (§11). Health insurance companies are obliged to publish specific criteria for signing

contracts on their websites at least once every 9 months and they need to include both types of criteria mentioned by the act: 1) personnel and technical resources of the provider and 2) quality indicators (§4). Quality indicators are designed in order to cover several dimensions of health care quality: accessibility of health care, efficiency of resource utilization, efficiency and appropriateness of health care delivery, patient satisfaction and outcomes of health care.

Definition of indicators and acceptable deviation is published separately by governmental decree no.752/2004 Z.z, later updated by governmental decree no. 51/2009 Z.z. and methods of calculation is set by decree of the ministry of health care. See Table 1 for the structure of the list of quality indicators as defined by the governmental decree and table 2 for the list of indicators defined for inpatient care.

Figure 1. Example of a quality indicator as defined by governmental decree 51/2009 Z.z.

Provider type	Dimension of health care	Indicator name	Indicator description	Time frame	Indicator level and acceptable deviation	Data source
Inpatient care provider	Outcomes of health care	Hospital wide mortality	Numerator Denominator	1 year	Level 0: Value above 2 times standard deviation from the mean Level 1: Value within 2 times standard deviation from the mean Level 2: Value lower than 2 times standard deviation from the mean	health insurance companies from providers

Figure 2. List of official indicators for inpatient sector

List of indicators for inpatient care	
Hospital wide mortality	Readmission overal within 30 days
Mortality after percutaneous coronary intervention	Readmission - overal within 90 days
Mortality after femoral bone fracture	Re-operation
Mortality on AMI - acute admission (age 35 - 74 years)	Readmission on J45.0 (Pneumony)
Mortality after stroke	Decubity
Mortality after hip replacement surgery	Nozocomial infection
Mortality after surgery	Operability
Mortality after interventions	Patient satisfaction
Maternal mortality	

Judging by the legislative framework, it might not be immediately apparent why there is a problem with performance and quality assessment of health care providers in Slovakia: the list of indicators more or less corresponds to what constitutes a good practice abroad and health insurance companies are obliged by law to publish these indicators and use them for selective reimbursement and contracting. Unfortunately, there are three main problems that hamper usefulness of the official quality indicators: 1) methodology of calculation is absolutely insufficient for any

meaningful comparison of the outcomes of care across providers 2) problem of low-quality unreliable data and lack of detailed data in the sector of inpatient care.

2.1.2. Methodology

Each insurance company calculates the set of indicators separately and sends the calculated values to UDZS (The Healthcare Surveillance Authority). Results are published separately for each insurance company, no overall evaluation of the indicators is performed by UDZS. Furthermore, presentation of data on health insurance websites is such, that you cannot compare 2-3 providers against each other because no exact values are published. That means that the quality indicators are used primarily for contract negotiation, if at all. Usefulness for consumer is practically nil.

As the primary aim of governmental quality indicators is differentiation of health care providers for the purposes of health care purchasing, methodology is a huge problem. There is no standardization employed, not even for age and sex, to account for different risk structures of patient populations of individual health care providers. The methodology completely disregards any issues of statistical significance – no confidence intervals are calculated to account for uncertainty in the data. Problematic is also definition of deviation from accepted standard: 2 standard deviations from mean value is a huge amount of tolerance – it can identify only absolute outliers, most providers fall within the accepted band. Differentiation of health care providers is therefore insufficient. No summary scores are calculated, making overall comparison of health care quality very difficult.

2.1.3. Data problems

Problem of insufficient and unreliable data presents a major obstacle to any attempt at provider profiling in Slovakia. Quality indicators are based primarily on the claims data submitted by health care providers to the heath insurance company in order to receive reimbursement for the health services provided to the insurees. Nature of the insurance claims is standardized by the regulation of UDZS - *Methodological directive no. 9/5/2006 Electronic processing and reporting of health services by the health care providers*, last updated in July 2011. In case of inpatient care, type of information reported to the insurance company is determined heavily by the specific nature of the reimbursement process in Slovakia – inpatient care is reimbursed using case-based system. Hospital is paid per completed hospitalization of a patient

according to the specialization of the department and type of hospital regardless of such parameters as the length of stay, severity of the patient's condition or the type of procedures actually provided (operations, examinations,…). Administrative data based on insurance claims therefore does include only minimum information on what health services were supplied – most importantly, claims report form does not include specification of what type of operation was performed (with the exception of those operations that have to be monitored for the purposes of maintaining waiting lists), which drugs were administered and when or which examinations were provided. In other words, once the patient is admitted to the hospital, the purchaser looses track of any and all treatment provided to the patient other than getting rough information about his movement between hospital departments. Consequently, we cannot calculate any process indicators to monitor adherence to evidence-based clinical guidelines, only the outcome indicators with limited risk-adjustment. See Appendix 1 for a list of information that should be included in the claims data (but many of the information not essential for reimbursement are not reported consistently).

Even more troublesome is the quality (or lack thereof) of those information about patient that are being reported to the insurance companies – namely diagnosis and comorbidities, both of which are fundamental for risk-adjustment. Because hospitals are not being reimbursed for case-mix of treated patients that would reflect severity of the patient's conditions, there is no motivation to invest any effort into reporting such data as co-morbidities at the admission or even major diagnosis. Similarly, there is no incentive to report data on patient safety and incidence of adverse events as these information are not related to reimbursement. Patient safety indicators, such as decubity or nosocomial infections which are usually essential part of any health care quality measurement, are therefore very unreliable in Slovakia. To illustrate the point, a large minority of hospitals does not even report on the mandatory adverse events reporting to UDZS or reported a total of zero adverse events [UDZS, 2011: pp. 39-41].

Another problem is caused by the common practice of splitting hospitalizations into several ones in order to increase reimbursement by the insurance company as hospital is paid per one case hospitalized regardless of the length of stay. Naturally, it makes more sense to record 3 separate hospitalizations, each lasting 2 days rather than 1 six day hospitalization. In order to have the claim approved by the insurance company, the patient is either moved to another department within the same hospital or another option is agreement between two hospitals – smaller hospital A prepares the patient fot the operation, larger hospital B then operates on the patient and hospital A then admits the patient again for post-operative care. Consequently it becomes rather

difficult to measure such seemingly simple indicators as length of stay or when measuring mortality, smaller hospital A might be misclassified as underperformer because its records show unexplained mortality – patients who die following operation in hospital B.

It will be the aim of my thesis to search for ways of how to deal with these problems and gain the best possible estimate of providers‘ quality given the data available at the moment and how it could evolve in light of envisioned introduction of DRG payment mechanism.

3. Methodological considerations

Health care providers‘ profiling is becoming increasingly common, often tied in with various reimbursement mechanisms (such as pay-for-performance), yet evidence on reliability of hospital profiling is very unsettling (for details see section 1.2.4.). Attention therefore must be paid to careful design of methodology in order to ensure robust results. In the following section I will briefly review main methodological considerations that need to be taken into account when designing hospital profiling with particular focus on mortality as a measure of quality. For a good summary of methodological issues in hospital profiling see Shahian et al. (2012).

3.1. Choice of indicators

We can distinguish three types of quality indicators as defined by Avenis Donabedian [Donabedian (1980)] – structure, process and outcome measures. Examples of structural measures include implementation of quality and review processes, nursing ratios and use of IT and other advance technology. Commonly used structure indicator is volume of procedures performed in the institution or per doctor. There is strong evidence that sufficient experience of doctors and nurses can be only achieved if a minimum volume of procedures is performed regularly – a study by Dr. Foster Intelligence - an independent UK initiative that regularly publishes UK hospital guide found that „death rate is reduced from 13 percent to eight percent when hospitals are doing more than 35 AAA procedures a year[2]“ [Dr. Foster Intelligence, 2011].

Process indicators assume existence of clinical guidelines that reflect best practice knowledge of how the care should be delivered. Process measures are rather common as they are relatively easy to measure[3] – they reflect adherence to certain prescribed processes, such as administration of beta-blockers prior to surgery in order to reduce risk of myocardial infarction[4]. No risk-adjustment or complex statistical methods are required to filter away variability in outcomes not caused by quality of delivered care. However, validity of the indicator is heavily dependent on the reporting quality of the

[2] abdominal aortic aneurysm

[3] even as they are usually rather costly to develop, they necessitate development of clinical guidelines, often on the basis of clinican trials

[4] for more process indicators see for example AHRQ website, Care Quality Commission U.K. or NICE that develops process indicators particularly for ambulatory care

hospitals and even more problematic is the fact that process measures have a tendency to focus attention on few specific areas/problems rather than evaluating overall quality of care: „however, there is some concern that excessive emphasis on achieving compliance with process measures...might lead to unnecessary screening procedures or treatments, or that they might conflict with a physician's best judgment or patient preference" [Normand and Shahian, 2007].

In the end, however, it is the outcome measures that we really want to measure, usually defined in terms of mortality or patient safety indicators such as incidence of preventable complications of care. Application of this approach to quality measurement, however, necessitates very careful risk-adjustment to account for case-mix of patients treated by the individual hospitals as well as techniques that can filter away variation in outcomes caused by chance. The main areas of concern regarding methodology of measuring outcomes are following: data sources, reference population (exclusion criteria), choice of indicators and observational period, selection of factors for risk-adjustment and heterogeneity of hospitals caused among other things by clustering of patients (i.e. non-independence of observations) and various sample sizes of patient populations.

3.2. Data sources

Source and quality of data is a crucial precondition for any succefsful statistical analysis. Profiling is usually done on administrative data – most readily available are claims data (see section 2.1.1.2. for description of claims data in Slovakia). In certain cases, these data are supplemented by clinical databases which significantly improve predictive ability of the models (see for example Tu et al. (1997)). Regarding the most common source of data – claims data, major concern is that their quality is heavily influenced by coding conventions in particular hospitals and precision of their recording (particularly in terms of diagnosis and comorbidities) and the fact that providers might have strong incentives to code claims data in certain way if these have impact on their reimbursement – such as diagnosis upcoding in case of DRG payment mechanisms. We are very likely to observe similar behaviour if the risk-adjustment mechanism is published (in order to increase credibility of the profiling efforts) including diagnosis codes of diagnosis categories and comorbidities used in the prediction model. Coding conventions should also be taken into account when importing models developed in other health care systems. Diagnosis codes that define certain conditions might differ significantly from region to region and should

always be validated before using (such as in the case of commonly used Charlson comorbidity score to code for comorbidities that comes with a list of diagnosis categories and diagnosis codes).

3.3. Choice of indicators and observational period

Various approaches exists on how to define mortality indicators, the basic distinction being made between hospital-wide and diagnosis/procedure specific mortality. In order to increase homogeneity of evaluated cases, it is often recommended to focus on a particular condition or procedure [Shahian et al., 2012]. Quite popular is mortality following coronary artery bypass grafting (CABG) as there is strong evidence for the existence of link between quality of care and outcomes in terms of survival – it is a complex procedure, yet if done properly, survival rates can be high [Normand et al., 2007]. At the same time it is a rather common procedure ensuring that we have a sufficient number of cases per hospital to allow for statistical testing. The downside is that such specific indicators capture only a minority of patients treated by the hospitals and furthermore, these measures rarely work without including any clinical indicators such as ejection fraction value in case of CABG [Normand et al., 2007]. Many of the smaller hospitals might not perform such procedures at all. In Slovakia, situation is further complicated by the fact that claims data do not include procedure codes (with the exception of operations for which official waiting lists exist[5]).

Therefore, next level of mortality indicators is a little less specific – mortality in high-risk diagnosis, mortality in low-risk diagnosis, death after surgery or death after planned (as opposed to acute) hospitalization. These mid-level indicators are attempting to balance the trade-off between homogeneity of evaluated cases and ability to draw conclusions on the overall quality of the hospital, not only quality of one surgical team or hospital ward. For example Canadian Institute for Health Information calculates surgical hospital standardized mortality rates [CIHI (2012)].

Finally, there is the hospital-wide mortality that estimates mortality across all the patients hospitalized in the particular hospital. This is by far the most controversial indicator as the level of heterogeneity it has to deal with is substantially larger than in

[5] Unfortunately, my data did not include information about procedure codes at all, nevertheless, I attempted to identify operations that included implantation of artificial joints as special medical materials, such as artificial joints, are reported on the claims separately.

case of more specific indicators – somehow the comparison has to be made across different diagnosis and in many cases the link between short-term mortality and quality of care is not established or simply does not exist [Shahian et al., 2011]. Various steps can be taken in order to decrease heterogeneity of cases included in hospital-wide mortality including exclusion of certain types of hospital episodes (oncological or trauma patients) or criterion to include only those diagnosis that account for at least 80% of deaths in the hospital.

Yet another dimension that needs to be considered when selecting indicators is the time period for which mortality should be assessed. Commonly used are in-hospital mortality or 30-day mortality, occasionally even 90-day mortality. Difference is crucial as hospitals might differ in their post-acute care options: some might discharge the patient to other institutions or home while other have limited options of discharging patient (no nursing homes in the vicinity) and have to keep him in acute beds and others might have departments of palliative care on their own. Hospitals also vary in distribution of treated cases and consequently, differences might arise in the likelihood of in-hospital death as opposed to long-term death. Shahian also argues that with our increasing ability of keeping alive very severe cases, indicator of acute-care in-hospital mortality becomes obsolete. [Shahian et al., 2012] Furthermore, the risk of gaming is high once the methodology is published: hospitals would be motivated to transfer difficult risky cases to other hospitals or discharge them prematurely in order to avoid in-hospital death.

3.4. Exclusion criteria

Definition of episodes and patients that are included (i.e. excluded) in the profiling methodology can significantly affect results. For example, study by Shahian et al. (2010) cited differences in patient cohorts as one of the main reasons behind large differences in results by various hospital performance measurement methods. After reviewing approaches of the four methods, they found that only 1/5 of all episodes was included in the analysis by all four methods. Shahian observes that these differences in patient populations reflect willingness of the profilers to tolerate in their models different levels of uncertainty in the association between mortality and hospital quality.

Among the episodes that are often excluded from the analysis are cases designated as end of life care – patients who are admitted to the hospital to die. Other possible

exclusions are trauma cases, oncological patients, but also newborns, episodes with length of stay exceeding 1 year or cases with „do not resuscitate“ on admission.

Similarly, decision needs to be made regarding hospital eligibility – there are various types of hospitals including general acute care hospitals, cancer centres, children's hospitals, heart institutes, even nursing homes and hospices. Inclusion of exclusion of certain types of hospital might significantly change mortality rates and predictions from the mortality model. It is therefore common to define peer groups – hospitals that are in certain respects alike and can be meaningfully compared between each other. For example CIHI hospital profiling distinguishes between small, medium and large community hospitals based on a number of patient-level and hospital-level characteristics including teaching status, number of inpatient cases and inpatient days and complexity of case (weighted number of inpatient cases). Specialty institutions are excluded altogether [CIHI (2012)].

3.5. Risk-adjustment

Regardless of the choice of mortality criterion, risk-adjustment allowing for particular patient-mix in the hospitals is crucial. Hospitals that treat mostly uncomplicated appendectomies can hardly be compared with hospitals that specialize in treating abdominal aneurysms or with hospitals that treat appendectomies as well, but do so in a region populated by elderly who suffer from multitude of health problems (due to for example location of mines or factories in the past).

Selection of the variables to use in the model for risk-adjustment must be done very carefully in order not to sacrifice validity in favor of good predictive power of the model [Normand et al., 2007]. This might happen if we confuse comorbidities for complications of care, adjusting for the length of stay, using diagnosis and comorbidities coded at the discharge from the hospital rather than admission and even inclusion of socioeconomic indicators must be carefully considered. We might filter away differences caused by different quality of hospital care. Such would be the case should we include length of stay group to approximate severity of the case – length of stay can be influenced by the severity of the patient's condition just as easily as by complications of care caused by low quality of provided care (infections, pneumonia, etc.). Another example when risk-adjustment is not desirable is the address of the patient – while it is true that patients from certain regions might have lower/higher probability of dying, different mortalities in certain regions might be

caused by the hospitals themselves rather than by the patients – if there are only one or two hospitals in the region and all are low-quality, mortality rates in the region will be predictably higher.

Risk factors that are commonly used in the hospital-wide mortality modelling usually include several dimensions. Listed below are examples of indicators that are often used in hospital profiling efforts:

- **socio-demographic factors** including age, sex and address of the patient, some proxy for socioeconomic status (in the UK so-called Carstairs index of deprivation is used that includes four variables – low social class, lack of car ownership, male unemployment and overcrowding), disability status;
- **diagnosis categories and procedure codes** (decision needs to be made as to whether include diagnosis at the admission or at the discharge and diagnosis codes need to be recoded into some meaningful aggregated diagnosis categories);
- **clinical variables** if available (BMI, blood glucose levels,...);
- **comorbidities** – conditions present at the time of admission that might affect results of care (most common are Charlson comorbidity and Elixhauser index);
- **severity of the episode**: ideally if severity of the episode can be estimated, for example episode grouper software such Thomson Reuters are able to construct severity scores. If such measures are not available, various proxies can be used including mode of admission to the hospital – emergency or other or transfer between hospitals;
- **past use of healthcare services** (past (emergency) hospitalizations, overall costs, costs for medications,...);
- **other relevant factors** such as dummy variable for surgical procedure, history of substance abuse of psychiatric problems or dummy variable for palliative care if these cases are not excluded.

3.6. Large datasets and modelling rare events

While it is true that more observations are usually considered good news due to the Central Limit Theorem, in certain cases large datasets might cause some problems.

Importantly, as the number of observations increases, all (or most) of the p-values are significant even as they are of little importance as they explain very little of the actual variability in the outcome of interest – in other words, some of the statistics lose their information value [Cerrito, 2010]. For example, diagnosis variable for malignant melanoma of skin DGN025 has mortality of 5.4% (more than twice as high as overall in-hospital mortality) and consequently has significant p-value of 0.000. However, only 25 patients with these diagnosis died out of overall number of deaths of 9 762, meaning that it explains all of a 0.26% of deaths, becoming virtually meaningless in the model. Similarly, Hosmer-Lemeshow test that is recommended as better measure of fit for logistic regression models than various R-squared statistics looses its usefulness with increasing sample size because it interprets even small differences between predicted and observed outcomes as significant deviation – no matter what you do, particularly when you wish for a reasonably parsimonious model, this test will reject the hypothesis of a good fit of your model [University of Strathclyde, 2012].

Second problem concerns modelling of rare events (events with lower than 10% probability). Precision of estimates from standard statistical models is decreasing as the error of the estimates increases and of course, size of the error if directly correlated to the number of events (among others). We must be therefore concerned about stability of the estimates when these are used for predictions on independent data. Cross-validation is essential in these instances for verification of stability of our estimates, but in cases of truly rare events it is recommended to use cross-validation not only for verification but also for derivation of the estimates themselves: final estimates are averages calculated across repeated sub-samples of data. A rule of thumb sometimes used to avoid these problems is a requirement of having a minimum of 10 events per variable [Peduzzi et al., 1996][6].

Another issue regarding rare events concerns usefulness of certain statistical measures. For example, commonly used measure of model accuracy for logistic regression – percentage of cases that were classified correctly becomes practically meaningless when dealing with rare events [Cerrito, 2010]. Model-free predictions without any explanatory variables yield a very high accuracy rate of 97.49% for in-hospital mortality - all the episodes with no death outcome are automatically classified correctly and as the mortality is only 2.5%, remaining (100-2.5%) of

[6] However, even if number of events is low, this might not necessarily be a problem - as long as variance of the data is limited. For example if we are modeling annual hospitals costs and we have a dummy variable of women over 90 with only 9 cases and each one of them has costs between 1000-1005 EUR, estimates will be stable.

observations are classified correctly. In this case, even if we compare the overall accuracy rate of the final risk-adjustment model with the accuracy rate of the null model, results are not particularly encouraging because the remaining accuracy rate is increasing only very reluctantly. This low increase in the accuracy prediction ratio is caused by the fact that almost none of the actual mortality is predicted in the model – the risk of adverse outcomes seems underestimated by this methodology. Therefore, we need to test for a presence of bias in the model: phenomenon called „model compression“ which occurs when adverse outcomes are underestimated and low-risk outcomes are overestimated [Schwartz et al., 2006].

3.7. Non-independence of episodes within hospitals

Traditional approach to hospital profiling is based on the assumption that there is no correlation between episodes at the level of providers – episodes are randomly distributed and independent. Multilevel models, also called mixed effects models are able to deal with this issue of clustered observation because they allow for dependence in outcomes within hospitals. These models are suitable for situations when we expect that two patients randomly selected from one hospital are expected to be more alike than two individuals selected from different hospitals. And these "group" characteristics are likely to influence the outcomes. If the clustering within hospitals are ignored, standard errors of the regression tend to be underestimated and p-values might be appear incorrectly significant. [Steele, 2012]

If clustering of observation is our only concern there exist aletrnatives to multilevel modelling, for example marginal models are sometimes used. However, if between-group variance in itself is of interest, multilevel modelling is the best option because it produces estimates of between-hospital variance, which is exactly what we are interested in hospital profiling [Steele, 2012]. This variable is usually denoted as τ^2 and is called random effects – total between-hospital variation beyond chance [Lingsma et al, 2010]. Interesting property of these random effects is the fact that their estimates are "shrunk" towards the mean [Normand et al, 2007]. This is beneficial for dealing with the issue of uneven sample sizes because groups with too few observations are drawing on the mean values. At the same these models also deal with the issue of statistical uncertainty as is obvious from the definition of random effects – "between-hospital variation beyond chance".

4. Data description

Original raw data consisted of three different datasets – data on all hospitalizations in the years 2010, 2011 and 1st half of the year 2012, data from medical claims of all the hospitalized patients prior to 12 months from their date of admission to the hospital and date of death, if applicable, of all the hospitalized patients. Furthermore, data on hospitalizations were originally used for different purposes and necessitated a good deal of cleaning and recoding since they consisted of thre separate datasets, each with a different structure and selection of variables. In the first part of this section I will briefly summarize major problems I encountered and decisions I made in order to prepare the final dataset for the analysis. In the second part I will describe all the variables that were used in the analysis.

4.1. Preparation of data

The process of data cleaning and recoding was following: 1) cleaning and recoding variables relevant for the purposes of creating "super-spells" of hospitalization cases which refers to continuous hospitalizations at one hospital regardless of number of transfers 2) cleaning duplicate and overlapping hospitalizations 3) recoding data to create hospital "episodes" 4) recoding of variables to reflect super-spells rather than ward-level hospitalizations and creation of new variables based on patient ID and date of admission to the hospital and 5) finally checking for any remaining problems and inconsistencies in the data.

4.1.1. Cleaning data

Duplicate and overlapping observations were identified and dealt with. In the first round, duplicate observations for the same patient, hospital, specialty of the hospital ward and date of admission and discharge were removed (these duplicate observations caused by merging two sources of data to create datasets – HOSPICOM and claims data, but also when additional charges in addition to hospital case were submitted, such as in case of special medical material, duplicate records might have been created in case of incorrect coding by the provider). Other duplicates were

caused by merging three separate datasets of hospitalization: a hospital case crossing over December, 31st 2011 was recorded in two datasets 2010 and 2011 (in case of 2010 dataset, date of discharge was imputed at December, 31st 2011 for all unfinished hospitalizations). Some spells have the same date of admission and same ID of the patient – this is not valid unless transfer occurred within 24 hours (recoded as transfer). In case of overlapping spells I kept the one with longer length of stay if possible, but several cases occurred with exactly the same length of stay – in these cases one of the overlapping episodes was always on the unit of intensive care and I arbitrarily dropped this one[7]. Newborns[8] were removed from the data early on as they have same identification number as the mother, thus creating overlapping and simultaneous hospitalizations. Hospitalizations with missing dates were dropped (primarily psychiatric patients) and in obvious cases mistakes in dates were corrected (for example 01feb2049 in the dataset 2011 was recoded into 01feb2011). Clean dataset contains 476 370 records.

After cleaning data and recoding the spells into episodes (units of observation described in the next section), I run some logical controls on the data in order to detect any remaining problems. These included for example distribution of age and sex for diagnosis of spontaneous birth (O8)[9], sex of the patients treated at OB/GYN department, distribution of age of patients treated at geriatric ward (56-102) and sex of the patients with the diagnosis of neoplasm of prostate (C61)[10]. Number of cases transferred to and from the hospitals do not match after the end of the recode (52 more transfers to the hospital than from the hospital) because, as will be explained in the next section, I discarded all the hospitalizations shorter than 24hours. If the transfer occurred within 24 hours of admission, the record of the transferring hospital was dropped.

[7] This is probably caused by the reimbursement mechanism – spell at the intensive care unit is not reimbursed in case of in-hospital transfers. This might have led to some incorrect records in the claims data.

[8] Information about the newborns is coded separately in the claims submitted by the provider. These, however, did not seem complete as there was a sizeable number of hospitalizations at the neonatology department (code 051, 194 or 203) that were missing the code for newborn. I tabulated these records with the age of the patients calculated on the basis of the ID and found that out of the 3,212 suspicious case 223 cases had 0 age indicating that the new ID was already assigned to the newborn and 2,989 cases had age between 15-47 indicating that we are indeed dealing with newborns recorded under the mother's ID.

[9] I detected 1 error – woman of 65 years with the diagnosis of birth at the surgical department. Diagnosis was recoded from the discharge data.

[10] 1 woman - diagnosis was recoded from the discharge data

4.1.2. Definition of the unit of analysis – "hospital episode" and exclusion criteria

Dataset on hospitalizations includes all the ended hospitalizations in the period 1.1.2010 – 30.6.2012[11], with the exception of one-day surgery procedures that are not considered. In the original data, one hospitalization refers to the hospitalization on one hospital ward – in case of in-hospital transfers, separate hospitalizations are recorded. Data were recoded in such a way as to define one hospital case in terms of overall stay in the hospital rather than stay in a single hospital ward – so-called "hospital episode" (otherwise we would not be able to evaluate mortality of course). Recode of separate hospitalizations into episodes was done via comparisons of date of discharge and date of admissions of separate hospitalization for patients with the same ID. Numerous variables then had to be recoded to reflect change in the units of analysis, where relevant value from the first hospital admission was used (for example in the case of dummy variable for planned hospitalization or in case of major admitting diagnosis), in other cases if the indicator appeared anytime during the super-spell it was recorded (such as in the variable indicating surgical procedure or long-term care).

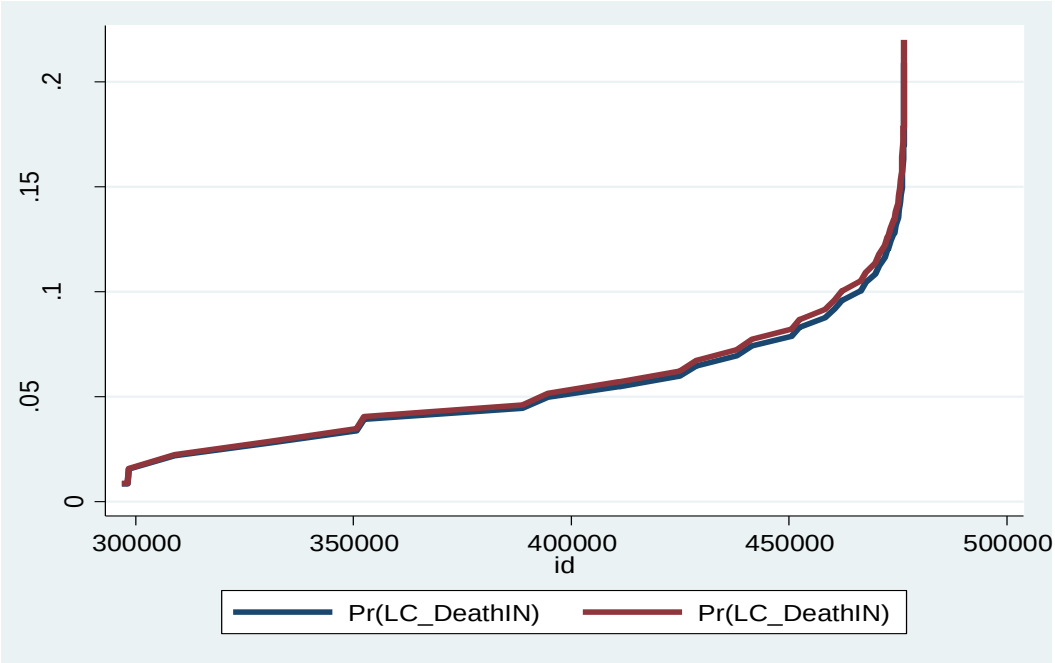
Transfers between hospitals are considered as separate admissions, which is somehow problematic when predicting death given that the outcome in transfers from the hospital is uniformly survival. Not only is this a problem for hospital comparisons as a hospital that transfer all the risky cases elsewhere would end up as the best pick (predicted probability of mortality non-zero, actual outcome 0), but the presence of these cases might affect coefficients of the predictive model. As an example, imagine a condition with 100% mortality when all of the cases are transferred for a specialized treatment from one hospital to another where they eventually die – predicted mortality for the condition immediately drops to 50% which is not a problem in itself but it is unfair to the receiving institution (predicted mortality of 50%, observed mortality of 100%) and undeservedly advantageous to the dispatching hospital. Figure 3 demonstrates this issue: model with only Charlson comorbidity score and diagnosis category of I21 (Acute myocardial infarction – one of the most frequent diagnosis of between-hospital transfers) yields lower probability of dying than identical model with the exclusion of cases transferred from the hospital (red line) – in other words predicting in-hospital mortality with episodes that were

[11] hospitalization that started before 1.1.2010 but ended after 1.1.2010 are also included, but hospitalizations that started before 30.6.2012 but ended after that date are not

transferred from one hospital lowers predicted probability of dying, disadvantaging receiving institutions.

Figure 3. Impact of excluding discharged transfers to another hospital from analysis.

Red line – probability of death after exclusion of transferred cases, blue line – probability of death for data including transferred cases. Receiving institutions is disadvantages if transferred cases are included (lower probability of dying).



One solution to the problem is to consider 30-day or 90-day mortality – death would be then assigned to both dispatching and receiving institution (that might cause exactly the same problem we covered in the previous paragraph – inflating probability of dying as one patient accounts for 2 deaths – 200% probability of dying, presenting unfair advantage to the receiving institution)[12]. **Therefore, I decided to exclude episodes that were transferred from one hospital to another** (episode in the dispatching institution was dropped, episode in the receiving institution remained)**.** This is not a perfect solution, percentage of cases transferred from the hospital is on average 3.5%, with the maximum of 12.6% cases ending in transfer. Taken altogether, I dropped 16 544 episodes.

Newborns were excluded from the data because mortality in this group is extremely low, 64 deaths out of 46 836 cases and majority of these cases are caused

[12] Dr Foster deals with this issue by creating so-called "super-spells" – hospital cases linked via between-hospital transfers. In other words the final outcome is assigned to both dispatching and receiving institution, thereby still looking at in-hospital mortality. However the problem why we did not opt for this solution remains the same as in case of 30-day mortality.

by prenatal problems that cannot by influenced by the hospital (such as low-birth weight). Furthermore, newborns cannot be viewed as a typical case for the hospitals as they were not admitted for suffering from a medical condition. Therefore, I will keep mothers in the dataset, but exclude newborns. Another option would be to introduce a dummy variable for a newborn, but the traditional models do not performs so well in case of extremely rare events – such as a death of a newborn (see section 3.2.4. for details).

In order to achieve best results possible given the fact that I am analyzing hospital-wide mortality, I focused exclusively on care provided in acute care institutions. The original data covered all types of institutions including nursing homes, palliative care institutions, sanatoriums and specialty hospitals such as as cancer centres, children's hospitals and heart institutes. All of these are excluded, **I analyzed only hospital episodes admitted to acute care institutions – university hospitals and general hospitals.** Reason behind this decision is the suspected existence of major differences in patient mix in different types of hospitals that can not be easily accounted for by the risk-adjustment model. Mortality in palliative care homes is 65.6% as compared to 0% mortality in sanatoriums treating lung diseases as compared to 0.5-6.3% unadjusted mortality in university and general hospitals. In case of specialty hospitals, problem arises as these institutions often treat rare diagnosis or very complicated cases that are not covered by the risk-adjustment categories (as I require certain minimum number of deaths and significant mortality for the diagnosis to be included in the risk-adjustment model). I will loose 73 362 episodes due to this restriction.

Similarly given that I am interested only in acute care, **hospitalizations longer than 365 days were excluded**. **Hospitalizations shorter than 24 hours were also dropped** because I reason that hospital had very little chance to actually make a difference in patients who died the same day they were admitted – these were probably were severe cases and it would not be fair to penalize hospitals for loosing the patient when they never had a chance of saving him in the first place. Similarly, we might consider excluding patients admitted with trauma or oncological diagnosis in order to maximize the link between the quality of the hospital and measures outcome – mortality.

Final number of episodes entering the analysis is 389 577 (523 206) – 75% of all the possible cases. Final dataset prepared for the analysis consists of 389 577 cases and 253 919 individual patients, average duration of one hospitalization was 7,5 days and median hospital stay was 5 days. All following analysis is performed on

derivation sample only – 75% from the sample dataset selected randomly. Remaining 25% will be used for validation of the final model (for details see section on methodology).

4.2. Descriptions of variables

In this section I will describe data and variables used for the analysis and perform some preliminary analysis of dependencies in data and transformation of data where applicable. First, I will focus on various mortality indicators that will be considered and in the second part I will examine explanatory variables – risk factors that will be considered in the mortality prediction model. All the analysis is performed on derivation sample only.

4.2.1.1. Hospitals

Data include episodes from 71 acute care hospitals, of which 61 will be included in the hospital profiling based on mortality. 10 hospitals will be excluded as the total number of treated episodes is lower than 100 over the period of 12 months (2011.07 – 2012.06) in order to ensure statistical significance of my results. For the purposes of developing mortality prediction model all the data are considered.

4.2.2. Dependent variables - mortality

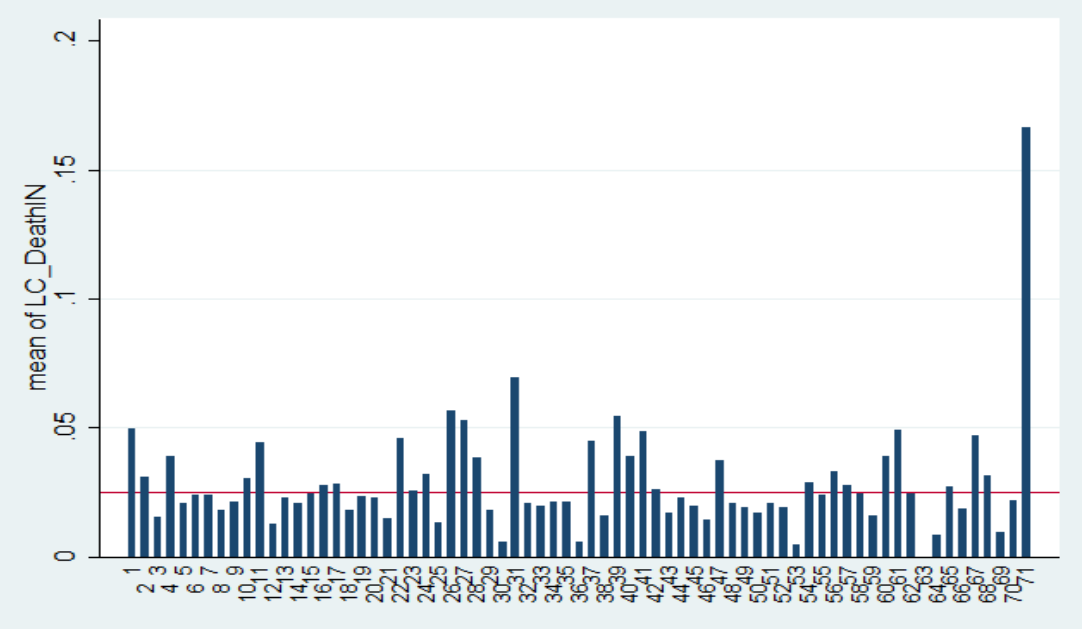
4.2.2.1. In-hospital all cause mortality

This variable captures mortality in the broadest sense - I introduced no further restrictions on diagnosis other than what was already described in the previous sections, although some authors recommend excluding for example oncological patients or trauma patients in order to eliminate cases where link between short-term mortality and quality of care is questionable.

What we are primarily interested if we want to profile hospitals on the basis of mortality is variability in mortality between hospitals. Figure 4 illustrates that **variability in hospital-wide mortality indeed exists and ranges from 0.5% to 7% in acute care hospitals**. **Overall mortality in the acute care hospitals is 2.49%** - that is 7 254 deaths.

Figure 4. In-hospital mortality by hospitals.

Red line represents overall in-hospital mortality



Data on deaths of the patients came from two sources – claims data where the providers have to record method of discharge from the hospital ward (transfer to other ward in the same hospital, to other hospital, release home, death) and data of the deaths of all the hospitalized patients from the records of insurance company. The match between the two datasets was far from perfect, usually I gave preference to the data from patient records of the insurance company, however in certain cases I trusted claims data instead – it takes some time to update records of the deceased patients in the records and therefore in cases of patients who died after 1jan2012 I used claims records where those indicated death and in other cases when mismatch existed I checked whether the patient received any healthcare after the alleged date of death. I used patient records where no other clue was available.

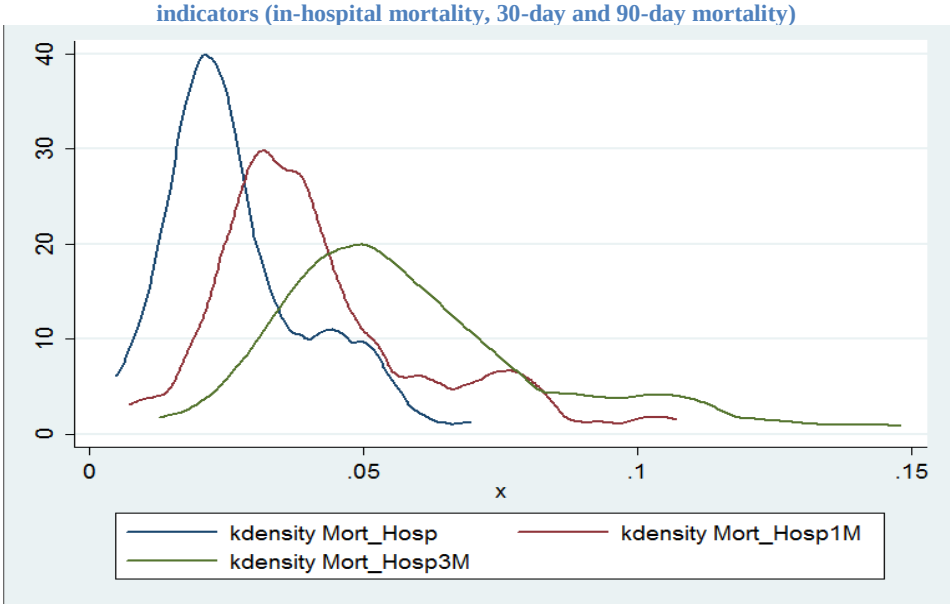
4.2.2.2. 30- and 90-day mortality

In case of 30-day and 90-day mortality, each hospital is assigned the death regardless of where it occurs – the hospital itself, other hospital or home. One death can therefore be attributed to two hospitals. As already mentioned, reasoning behind using longer observational periods is to level the field by disregarding discharge practices of different hospitals and to acknowledge that medicine is able to prolong life sufficiently to render indicator of short-term mortality virtually meaningless. On

the other hand, we need to weight these arguments against the risk of assigning unrelated deaths to the hospital in question (for example in case of between hospital transfers, if the receiving hospital screwed up discharging hospital will be unfairly penalized).

30-day mortality rate is higher than in-hospital measure, up to **3.92%** on average. **Variability between hospitals increased to 0.7%-10.7%**. Average **90-day mortality increases to 5.64% with distribution between 1.3% and 14.8%.** See Figure 5 for comparison of differences in variability of mortality rate between hospitals for the three mortality indicators.

Figure 5. Comparison of distribution (kernel density) of mortality between hospitals for three mortality indicators (in-hospital mortality, 30-day and 90-day mortality)



4.2.2.3. Mortality in high-risk diagnosis

This mortality indicator focuses on the quality of care provided in very risky episodes: it evaluates outcomes for a selected number of diagnosis categories – those that have mortality rate over 10% (for list of high-risk diagnosis categories see Appendix 2). By doing so, this indicator should be able to focus on a more homogenous patient population while ensuring sufficient number of cases for hospital profiling. Constructed high-risk indicator includes 12 diagnosis categories with the **overall mortality of 13.5%**, the most "deadly" diagnosis being cardiac arrest (I46 – mortality of 76.6%) and respiratory failure (J96 – mortality of 35.9%). Altogether 23 710 episodes are classified as high risk and **55 hospitals have over 100 episodes classified as high risk, mortality between hospitals varies from 4.9% to 26.6%.**

High-risk diagnosis categories will be also used for risk-adjustment in hospital-wide indicators of mortality.

4.2.2.4. Mortality in low-risk categories

Focus of this indicator is on so-called "never events" – deaths that should not have occurred. This might be seen as a proxy for safety indicators – while complications of care caused by the providers of care themselves are not reported consistently or reliably, mortality (but also overly long hospitalizations) in low-risk diagnosis can be detected and they indicate potentially serious problems in the hospital. For the identification of low-risk diagnosis, I used list of diagnosis codes published by Dr. Foster Intelligence – an independent UK initiative for health care providers' profiling (see Appendix 3).

66 hospitals are eligible for comparison (number of episodes with low-risk diagnosis >100) and hospital mortality ranges from 0% (10 hospitals) to 1.7%. There are very few cases of deaths - with overall mortality of 0.3% there are only 331 cases of low-risk deaths. Therefore, running a statistical analysis similar to the other mortality indicator is not viable. Instead, I will use low-risk mortality for validation of my exercise in hospital profiling – the test will be to see whether hospitals with above-average mortality in low-risk diagnosis were correctly identified by methodology. Should there be significance discord, I would have to re-evaluate my approach.

Low-risk diagnosis categories will be also used for risk-adjustment in hospital-wide mortality indicators.

4.2.2.5. Mortality after planned hospitalizations

Planned hospitalizations are those that are not acute and therefore do not require immediate admission to the hospital. Advantage of this indicator is the value for the consumers – in acute cases, patients do not really have the liberty of choosing hospital to be treated in as they needs to be admitted immediately. When the hospitalization can be postponed, the patient can choose high-quality hospital and an indicator of how well the hospital treats non-acute cases might be useful in these cases. 15.46% of all episodes are planned hospitalizations.

Mortality for planned hospitalizations is rather low at 0.92%. This low mortality will probably present problem in mortality modelling – we are virtually attempting to

model and predict a "never event". **Hospital mortality ranges from 0% up to 24%** (existence of obvious outliers as the next highest mortality is 5.4%).

4.2.2.6. Mortality after surgery

Performace of a surgery poses serious risks for the patients and high mortality after surgical procedures may indicate one of two problems – complications developed during the surgery or as a direct consequence of surgery or asking whether the surgery should have been performed in the first place [Hospital guide 2011]. **Mortality is again rather low at 1.4%** and reaches values between 0% and 6.2% for individual hospitals. 26.6% of episodes included a surgery.

4.2.2.7. Mortality after implantation of artificial joint

This mortality indicator is for illustrative purposes only as the data I had were insufficient for identification of specific procedures. Generally, providers do not report specific procedures to the insurance company, with the exception of diagnosis for which official waiting lists have to be maintained by the insurance companies. While I did not have codes for these procedures available, I had access to all the health services provided during the hospitalization (from claims data), including additional special medical material needed during the hospitalization that is not covered by the case rate payment. One of the procedures that require use of individually reimbursed medical material is implantation of artificial joint (hip or knee). I identified all the episodes that involved provision of artificial joints based on the product codes of the implants[13] (see Appendix 4).

Only 20 hospitals performed more than 50 of the procedures, mortality rate after the implantation of artificial joint is 0.2% (which is consistent with findings that post-operative mortality after joint replacements is generally low). However, such low numbers and rare occurrence cannot be adequately modeled or predicted by standard statistical techniques and therefore, this mortality indicator will not be considered further in the analysis.

4.3. Risk factors – explanatory variables

Figure 6 summarizes all the potential risk factors that will be considered in the mortality modelling as risk-adjustment variables.

[13] Yet I probably failed to capture all the procedures: in 2011 alone I identified 1 163 operations, while official data of the insurance company state number of joint replacements at 1472 per annum.

Figure 6. Description of candidate variables for risk-adjustment (derivation sample only)

Includes variable name, description including values, prevalence, number of non-survivors, % of non-survivors out of all cases classified, % of non-survivors out of all deaths and chi-square test testing for existence of differences between categorical variables (number of deaths and variable in question)

Descriptive statistics - potential explanatory variables derivation sample				
Variable name	**Description***	**Prevalence** (% classified/all episodes)	**Non-survivors** No deaths (% deaths/all classified; % deaths/all deaths)	**Chi-square test** (S - significant at 95%)
Demographic variables				
LP_Sex - female				
female		58.3%	3381 (1.98%; 1.16%)	S
male		41.7%	3873 (3.19%; 1.33%)	S
LP_Age*				
age0	0-9			S
age1	age 10-19	7.55%	10 (0.05%; 0.14%)	S
age2	age 20-29	12.92%	48 (0.13% ; 0.66 %)	S
age3	age 30-39	12.88%	141(0.38%; 1.94%)	S
age4	age 40-49	9.32 %	363 (1.34%; 5.00 %)	S
age5	age 50-59	14.42%	1100(2.65%; 15.30%)	S
age6	age 60-69	13.93%	1668 (4.12 %; 22.99%)	S
age7	age 70-79	10.60%	1821 (5.91%; 25.10%)	S
age8	age 80-89	5.26 %	1768 (11.56%; 24.37%)	S
age9	age >90	0.53%	293 (18.87%; 4.04%)	S
Diagnosis categories**				
DGN_LR	low risk diagnosis group	38.69%	331 (0.29%; 4.56%)	S
DGN_MR	medium risk diagnosis group	7.53 %	1329 (6.07%; 18.32%)	S
DGN_HR	high risk diagnosis group	8.15%	3199 (13.49%; 44.10%)	S
DGN01	Malignant neoplasms of digestive organs	1.19%	358 (10.33%; 4.94%)	S
DGN02	Malignant neoplasms of respiratory and intrathoracic organs	0.74%	222 (10.59%; 3.06%)	S
DGN03	Other neoplasms	1.85%	339 (6.31%; 4.67%)	S
DGN04	Endocrine, nutritional and metabolic diseases	1.67%	453 (9.30%; 6.24%)	S
DGN05	Ischemic heart disease	2.90%	354 (4.24%; 4.88%)	S
DGN06	Other heart diseases	1.96%	303 (5.33%; 4.18%)	S
DGN07	Heart failure	1.23%	494 (13.71%; 6.81%)	S
DGN08	Cerebral infarction	1.09%	316 (10.07%; 4.36%)	S
DGN09	Other cerebrovascular disease	1.48%	521 (12.17%; 7.18%)	S
DGN10	Influenza and pneumonia	2.42%	312 (4.40%; 4.30%)	S
DGN11	Other diseases of respiratory system	3.57%	490 (4.72%; 6.75%)	S
DGN12	Diseases of liver	0.73%	256 (12.08%; 3.53%)	S
DGN13	Other diseases of digestive system	5.23%	560 (3.68%; 7.72%)	S
DGN14	General symptoms and signs	1.06%	69 (2.22%; 0.95%)	not S, added to DGN17
DGN15	Other symptoms involving specific system	1.12%	96 (2.93%; 1.32%)	not S, added to DGN17
DGN16	Injuries	2.59%	311 (4.13%; 4.29%)	S
DGN17	Other	6.36%	622 (3.36%; 8.57%)	S
Comorbidities				
Charlson Comorbidity Score*				
ch0	Charlson score 0 - no comorbidities	64.72%	1279 (0.68%; 17.63%)	S
ch1	Charlson score 1-10	17.65%	1667 (3.25%; 22.98%)	S
ch2	Charlson score 11-20	12.51%	2591 (7.12%; 35.72%)	S
ch3	Charlson score 21-30	3.83%	1149 (10.31%; 15.84%)	S
ch4	Charlson score 31-40	1.04%	438 (14.44%; 6.04%)	S
ch5	Charlson score >40	0.24%	130 (18.95%; 1.79%)	S

Descriptive statistics - potential explanatory variables derivation sample				
Variable name	**Description***	**Prevalence** (% classified/all episodes)	**Non-survivors** No deaths (% deaths/all classified; % deaths/all deaths)	**Chi-square test** (S - significant at 95%)
Severity of the episode				
LP_LTCare	Hospitalization at the ward of palliative or long-term care	4.78%	1722 (12.38%; 0.6%)	S
LC_TransferIN	In-hospital transfer of the patient between hospital wards	9.05%	2506 (9.52%; 0.9%)	S
LC_Surgery	surgical procedure performed during episode	26.62%	1105 (1.43%; 0.38%)	S
LC_TransferIN	admitted transfer	2.32%	638 (9.44%; 0.22%)	S
LC_EMERG2_Transp	emergency transfer to the hospital	19.67%	3664 (6.40%; 1,26%)	S
LC_EMERG3_Spec		26.73%	4204 (5.41% ; 1.45%)	S
LC_EMERG1_ARO	admission to the unit of intensive care within 24 hours of admission to the hospital	0.86%	1012 (40.27%; 0.35%)	S
LC_PlannedHosp	non-acute episode	15.26%	385 (0.87; 0.13%)	S
Past use of healthcare services				
LP_NoHosp*	Number of hospitalizations in the past year	28.38% (>0)	3680(4.37%; 1.27%)	S
==0	no past hosp.	71.06%	3574 (1.73%; 49.27%)	S
==1	1-3 past hosp	23,81%	2527 (3.65%; 34.84%)	S
==2	>3 past hosp	5.12 %	1153 (7.74%; 15.89%)	S
Past use of healthcare services				

** unless stated otherwise, dummy variable 0-1*

*** see Appendix for the diagnosis codes*

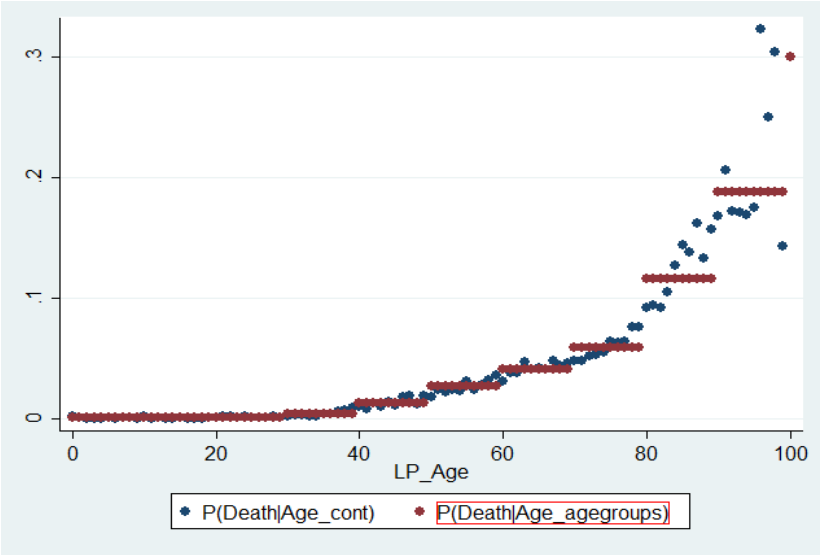
4.3.1.1. Age

Age of the patient at the day of admission to the hospital recoded from the patient identification number[14]. Graph in Figure 7 depicts probability of death for hospitalized patients with the rising age – dependence between mortality and age is obvious and increasing with age.

[14] Variable was recoded from the variable **LP_IDpatient** as the ID number is created according to the following formula: YYMMDD, where 1st and 2nd digits denote year of birth, the 3rd a 4th digits stand for month of the birth for the males and month of the birth+50 for females and 5th and 6th digits stand for day of birth. Difference between people born in 1912 and 2012 can be identified by number of characters in the ID - people born before 1954 have a 9-digit ID and people born later than 1953 have a 10digit ID. Age was then calculated as a difference between the date of admission to the hospital and date of birth. 108 observations were problematic as the values for month were higher than 62. I extracted the age variable from the 2012_Hospitalizations dataset that included (as the only one) also age variable, but that still left 83 observations with unidentifiable age and consequently also sex. I could have approximated the age as I has the year of birth however, I still would be missing data for sex. Therefore, I decided to drop these 83 observations. The new variable has values between 0 and 103.

Figure 7. Probability of death by age (mean probability of death per 1 year/age group).

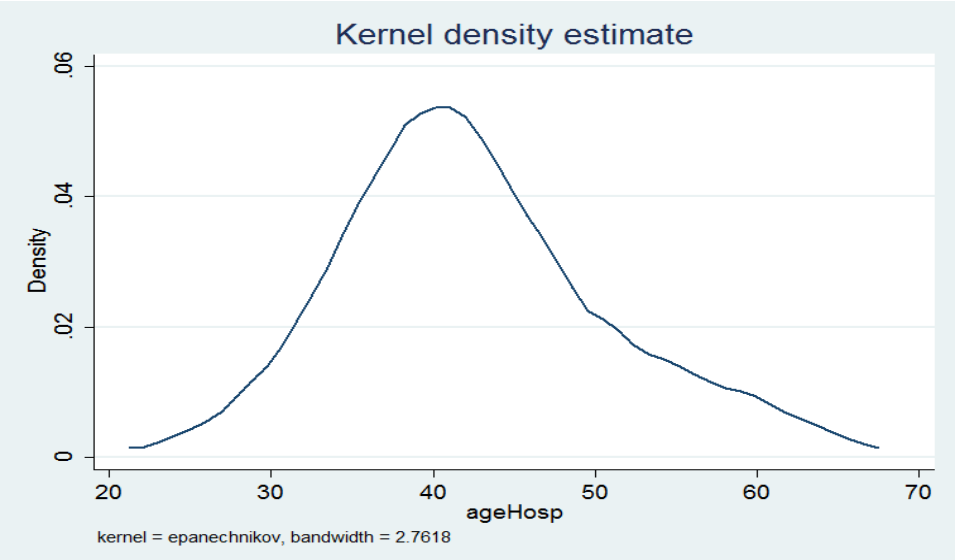
Blue dots – original continuous variable, red dots – transformed variable into 10 dummy variables (10-year bands)



Crucial question concerns fitting the variable into the model – how the variable should be transformed in order to perform well in the model. Usually, the age variable is recoded into 5-year (or 10-year) bands for two reasons – first, if majority of the variables are dummies with values of 0 and 1, variable with values up to 100 has a tendency to dominate the model – it might appear to be more important than it really is [Cerrito, 2010]. Second, age is rarely a linear variable – effect on the variable of interest (outcome) does not increase by the same value for 10-year old and 60-year old - instead in Figure 7 we can see that the effect of age on probability of dying is clearly non-linear – at higher age probability of dying with each additional year is increasingly much more rapidly than at young age. This exponential relationship can usually be fitted very well into logit models because as the probability is exponential, ln(p/1-p) is linear [Lunt, 2012]. However, further on in the analysis Box-Tidwell test of the linearity of the model indicated problem with the continuous age and therefore, I decided to recode age into 10 10-year bands categories.

What is equally important in terms of hospital profiling is different risk structure of individual hospitals regarding the risk adjustment variables. Figure 8 illustrates significant difference in age structure of treated episodes by hospitals: average age of patient population in a hospital is 42 years, with the minimum of 24 years and maximum of 65 years.

Figure 8. Kernel density of mean age of treated episodes by hospitals (continuous age)



4.3.1.2. Sex

Sex of the patient recoded from the ID of the patient (1 – man, 2 - women). Chances of dying if you are a male hospitalized in acute care hospital are higher by 63% compared to females. Lower mortality of females is partially caused by the presence of deliveries – they account for high number of episodes yet mortality is virtually non-existent (30 176 pregnancy related episodes and deliveries and no death). After excluding these cases, mortality of females rises moderately to .0229149 – which still makes for 40% higher mortality of males in the hospitals.

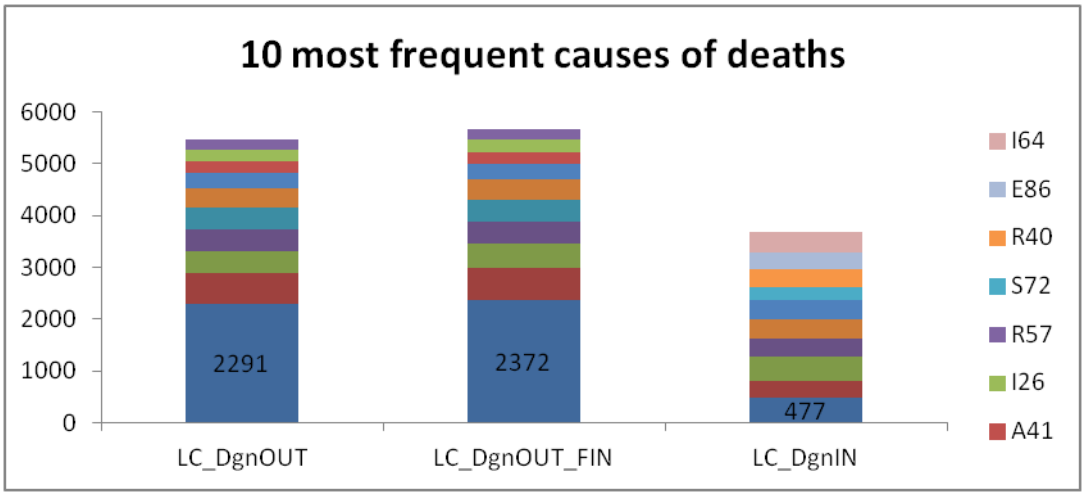
4.3.1.3. Diagnosis categories

Selection of variable

First of all, we need to decide which diagnosis code to use for the purposes of classification of the episode into the diagnosis. I considered several options – diagnosis recorded at the admission - LC_DgnIN, diagnosis recorded at the discharge from the first hospital ward - LC_DgnOUT and diagnosis at the discharge from the hospital - LC_DgnOUT_FIN. Differences in the diagnosis codes responsible for most deaths using these three variables were notable. When looking at the distribution of the diagnosis of these three variables, it becomes apparent that the two variables looking at the diagnosis at the discharge are much more concentrated: the first 10

most common causes of deaths account for about 50% of total deaths, while is amounts to only 23% in the diagnosis recorded at the admission (Figure 9). This is caused by the coding practice in the hospitals – when death occurs, physicians often code the immediate cause of death as the discharge diagnosis – which is heart failure (I50) in most cases (frequency of I50 as the diagnosis in fatalities increased by 80% in discharge variables in comparison to admission variables). Prevalence of I50 diagnosis is even higher when we consider LC_DgnOUT_FIN variable.

Figure 9. 10 diagnosis responsible for most in-hospital deaths – three candidate diagnosis variables



However, we are more interested in the „real“ cause of death, therefore LC_DgnIN seems as a better option. Furthermore, discharge diagnosis may reflect complications of care caused by the process of care itself rather than the medical problem necessitating admission to the hospital in the first place. Using the discharge variable is therefore not recommended for the purpose of ranking providers. Shortcoming of using LC_DgnIN variable is the fact that diagnosis at the admission is often unknown - the recording physician can only note the obvious cause of admission as accurate diagnosis needs further tests and examinations.

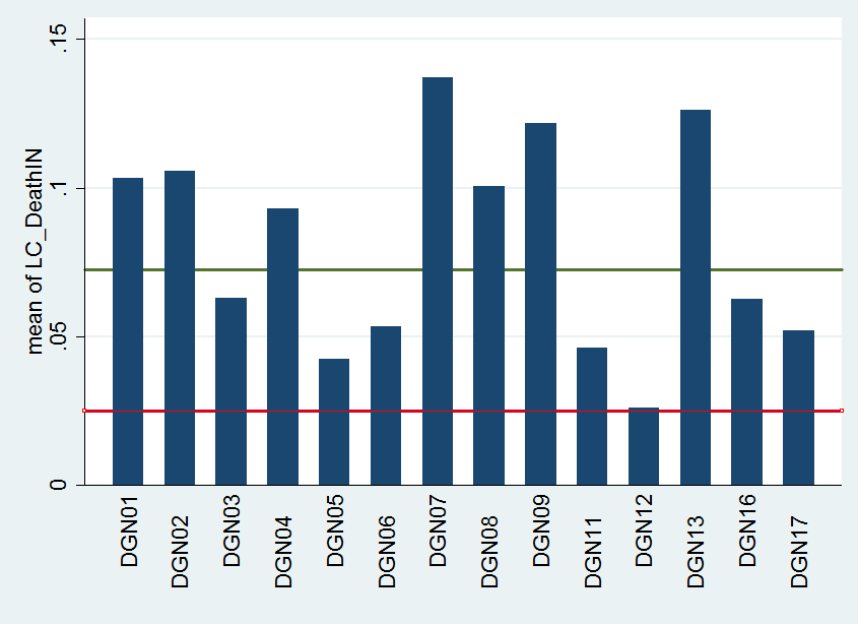
I opted for a compromise solution: by default I considered admitting diagnosis but in the cases when diagnosis recorded at the admission was non-specific (ICD10 codes starting with R - Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified) I used admitting diagnosis for the following transfer if applicable, if not I used discharge diagnosis. This modification involved 21 319 cases and while majority of the diagnosis remained unspecified even after the recode, we managed to specify at least some of the diagnosis – most frequently K30 – dyspepsia and J20.9 – Acute bronchitis, unspecified.

Deriving diagnosis categories

Diagnosis categories were constructed by taking all the 3-digits diagnosis codes and selecting those that account for 80% of all deaths in the hospitals. In the next step diseases were ordered by codes and grouped by physiological similarity (according to ICD10 – International classification of diseases) into 17 diagnosis categories (see Appendix 5 for a list of diagnosis categories and codes). Groups were formed gradually by merging low-count categories into larger groupings until a reasonable number of deaths was achieved in each diagnosis grouping. 2 of these categories were later merged with DGN17 – Other because chi-square test indicated that differences in mortality for those that code positive for the diagnosis category and those that do not are not significant. Overall**, I have 15 diagnosis groups, 35% of all episodes have assigned one of diagnosis categories, they account for 81% of in-hospital deaths and mortality in these categories is 5.8%** - that is twice the overall in-hospital mortality. Variance in the mortality by categories is substantial – see Figure 10.

Figure 10. Mortality by diagnosis categories.

Green line: average mortality in episodes classified in one of the diagnosis categories (7.25%), red line: overall average mortality (2.49%). See Appendix 2 for definition of diagnosis categories



High-risk diagnosis

Dummy variable that indicates whether the episode is classified as high-risk based on the diagnosis codes. Diagnosis with mortality over 10% were selected (see Appendix 2), including 23 710 episodes. Overall mortality with episodes in high-risk category

is 13.5% compared to mortality of 1.5% in the all the other episodes. Importantly**, risk structure of episodes for individual hospitals differs considerably: ratio of high-risk episodes on all the episodes treated differs from 2.3% up to 30.2%**.

Low-risk diagnosis

Dummy variable that indicates low-risk episodes on the basis of diagnosis codes. List of diagnosis codes was adopted from Dr Foster's methodology (see Appendix 3). 38.7% of all the episodes are categorized as low-risk – mortality for this group is extremely low at 0.29%.

4.3.1.4. Charlson comorbidity index

Charlson comorbidity index is one of the most well-known measures developed to capture patient comorbidities specifically for the purposes of predicting hospital mortality [Charlson et al., 1987]. It is often used in academic literature, not so much for the actual hospital rankings as it is a published index and therefore carries the danger of upcoding because relevant disease categories used for risk-adjustment are known to the hospitals that might adjust their coding practices accordingly [Cerrito, 2010]. The index is based on 17 diseases that have assigned weights based on their probability of death. The original index assigned weights from 1 to 6 (highest number assigned to single disease category - HIV). Due to possible issues of multicollinearity (since the index includes for example both diabetes and diabetes with complications) [Cerrito, 2010] and given the practical consideration of minimizing the number of dependent variables, the index is expressed as a single number[15]: all of the patient's comorbidities at the time of admission to the hospital are identified, assigned the weight and summed up to produce a single score. Before I proceed, a note of caution: this index is afflicted by the same shortcoming as most risk-adjustment indices – it captures only a limited number of diagnosis – severe but rare diagnosis are not captured [Cerrito, 2010]. Given a shortlist of disease categories the risk-adjustment may capture differences in coding practices rather than patient severity.

In my analysis, I used a modified Charlson index – first of all, I needed translation into ICD10 codes used in Slovakia and secondly, I used recalculated weights for disease categories as published by Clinical Indicators Team in the UK [NHS, 2012]

[15] alternative to Charlson index is Elixhauser index which computes dummy variable for each comorbidity condition

which range from 1 to 18 (maximum of 18 points for severe liver disease) (for detailed list of diseases, codes and weights see Appendix 6). The original index was calculated in 1987 on the US data and the mortality risks of individual diseases have changed since then – for example HIV is hardly the single most risky disease in terms of in-hospital mortality. Ideally, of course, the weights should be calculated to fit the particular population they are applied to – patterns of mortality might be different in the Slovakia than in the US. Furthermore, also the disease list should be reviewed to ensure that the 17 diseases included are the most significant predictors of patient severity. Last, but not least, the list should reflect local coding practices – if most of the codes on the claims data are 3-digit codes whereas the Charlson comorbidities are often defined by 5-digit diagnosis codes, the index will not work.

Crucial obstacle to construction Charlson score in my data is presented by the fact that coding for comorbidities in the claims data is not reliable (and using unreliable data is worse than using not data at all in this case) and was not available at all in my case. I resolved to construct Charlson index on the basis of historical data – I had access to all health care provided to hospitalized patients for duration of 12 months prior to their admission to the hospital. In the first step, all healthcare records with diagnosis code that corresponded to one of the seventeen Charlson diseases were selected. All records were assigned value of 1, hospitalizations were assigned value of 10 and pharmaceuticals, specialized medical material and specialized ambulatory care were assigned value of 2. Values were summed up across individuals and Charlson diseases and only those patients were assigned Charlson category who reached a minimum value of 10 in the category (i.e. 1 hospitalization with a Charlson disease code was adequate to assign to the patient corresponding comorbidity). This was an arbitrary decision and further analysis is probably a good idea to explore effects of choosing different methods on the Charlson scores. Furthermore, classification that would rely on consumption of specific healthcare services (procedures and medicines) rather than diagnosis codes should be explored (PCG – pharmaceutical cost groups).

Next step is the validation of the newly-constructed indicator. Most of all we need to verify correlation between the Charlson score and mortality - if the index is defined correctly we should observe increasing mortality as the Charlson score increases.

Regarding the 1st requirement Figure 11 demonstrates that mortality has an increasing trend, but fluctuates somewhat, primarily for higher scores. This is caused by very few observations in these categories (for example there is only one patient with the index of 83). I will therefore recode index into 6 categories – i.e. 5 dummy variables.

Figure 12 shows how well this transformation fits the original data in terms of probability of death. It is also obvious that mortality is increasing with recoded Charlson. In the final version of the Charlson index, probability of death of patients with Charlson score of 0 – no comorbidity is 0,67% and mortality of patients with index of 5 is 18,95%.

Figure 11. Mortality by Charlson Comorbidity index – original score.

Red dots: mean probability of death by Charlson index (original index ranging from 0 to 83). Blue dots: fitted probability of death calculated from the logit model: LC_DeathIN Charlson Index

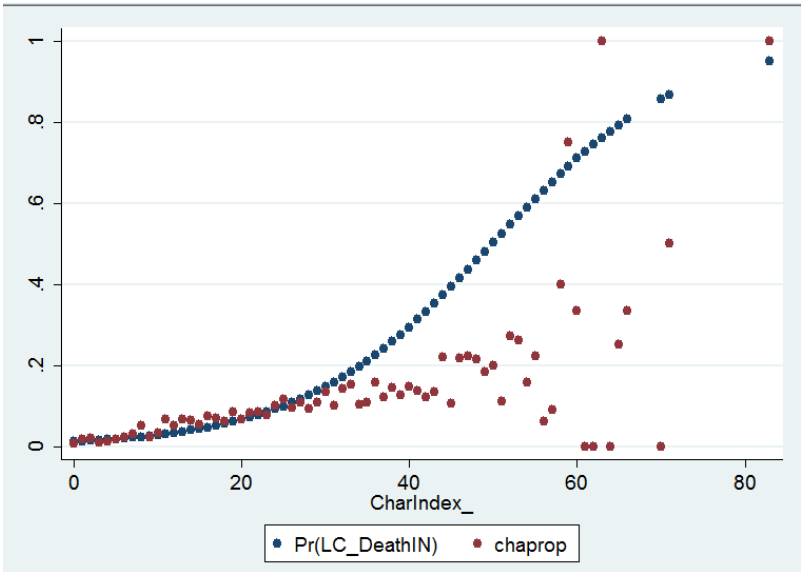
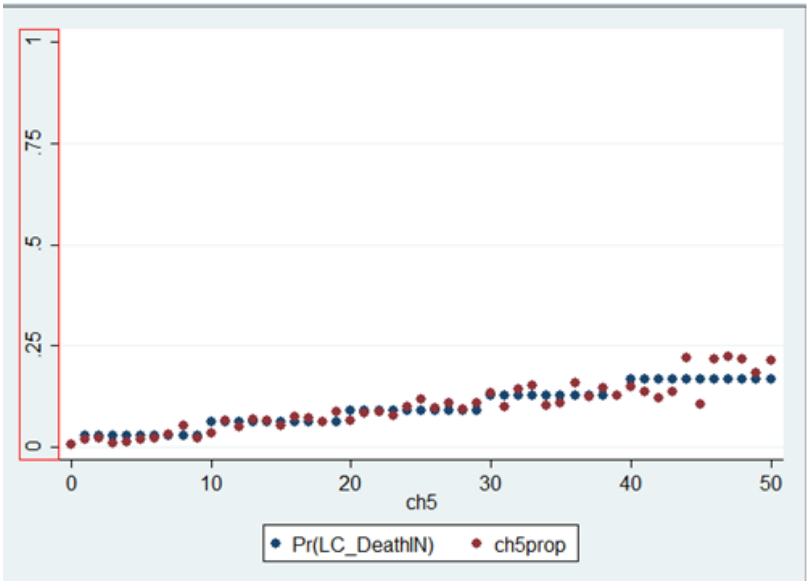


Figure 12. Mortality by Charlson comorbidity index – after transformation

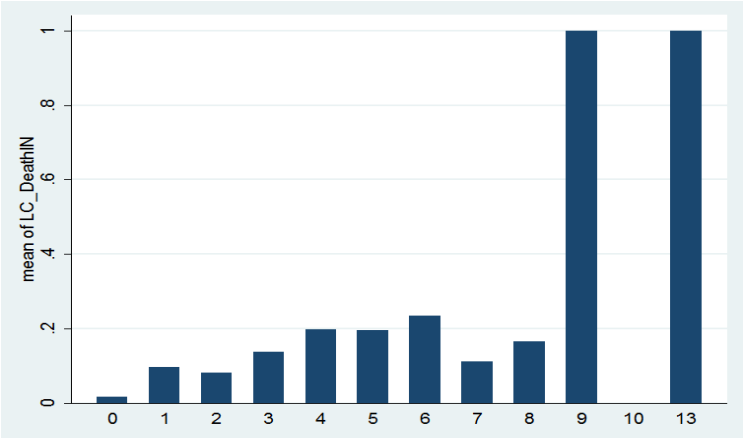
Charlson score is capped at 50 points and transformed into 6 dummy variables. Graph illustrates how well the fitted variable (blue) predicts actual observations (red).



4.3.1.5. Transfer between hospitals

Transfers between hospitals can be seen as partial indicator of severity of the patient – routine cases with no complications are not expected to be transferred. Patients that were transferred to other hospital are nearly four times as likely to die as patients that were treated in one hospital only even though it was already mentioned that some of the transfers are caused by the practice of rotating patient between hospitals in order to obtain several payments for the same case. The original variable that included actual number of transfers was not monotonous in relation to in-hospital mortality (see Figure 13) and therefore was recoded into a dummy variable: 0 – no transfer, 1 – 1 and more in-hospital transfers.

Figure 13: Mortality of patients by number of between-hospital transfers.



4.3.1.6. Admission to unit of intensive care within 24 hours of admission to the hospital

This dummy variable identifies those cases that were admitted to department of intensive care (ward specialty code 025) within one day of their admission to the hospital. Variable therefore identifies extremely severe cases. Restriction to identify only those cases transferred intensive care within 24 hours was introduced in order to prevent adjusting for complications caused by inappropriate care.

4.3.1.7. Emergency admission

Two candidate variables are proposed – admission after emergency transport[16] and admission to the hospital ward associated with emergency medicine (following

[16] type RZP and RLP

specialties: 174, 175, 176, 177, 184, 179, 180, 181). As in the previous indicator, these variables were designed in order to approximate severity of the episode. The two variables are highly correlated (Pearson correlation coefficient of 0.81) and therefore only one can be used in the risk-adjustment model, otherwise multicollinearity would be present. Some correlation exists also between both of these two candidate variables and admission to unit of intensive care (0.1 and 0.08), but certain amount of multicollinearity can be accommodated by the model. Nevertheless, multicollinearity diagnostics will be performed in order to make sure that multicollinearity is not a major problem.

4.3.1.8. Long-term care

Dummy variable that captures long-term and palliative care of the patient (specialty codes 620 – hospice care, 193 – hospice care, 334 - palliative care and 205 –long term care). While specialized palliative institutes are excluded from the analysis, some long-term and palliative care is provided also in the acute-care hospitals and for those cases I introduced dummy variable LT_LTcare. An episode is labelled as long-term care if the patient was previously hospitalized at a ward of a long-term care (hospital wards as defined in the first sentence). This variable is an indicator of long-term serious medical problems of the treated patient.

4.3.1.9. Number of hospitalizations in 12 months prior to admission to the hospital

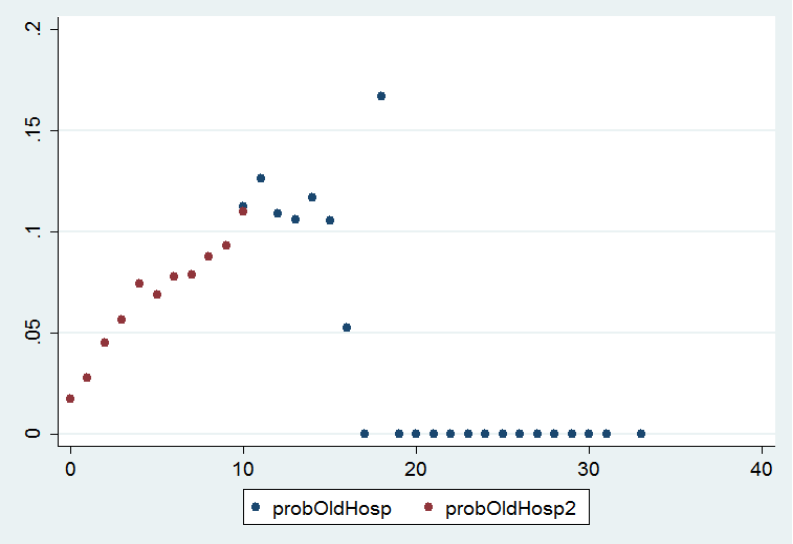
Past use of healthcare resources is also sometimes considered as a risk-adjustment variable in mortality prediction models, including overall costs, pharmaceutical costs, number of previous hospitalizations or number of previous emergency hospitalizations. In my analysis I constructed variable identifying number of previous hospitalizations in the past 12 months prior to the date of admission.

In order for this variable to be included in the model I need to verify existence of link between number of past hospitalizations and mortality. Figure 14 depicts mean mortality by number of hospitalizations – blue dots is the original variable (which is obviously problematic for episodes with over 10 past hospitalizations[17]) and red dots

[17] This might have been caused by the fact that I did not distinguish between separate episodes – which should have been done in order to obtain accuracy. Otherwise 5 hospitalizations might be in fact only 1 hospitalization including 4 transfers. Indeed, having over 10separate episodes in the past year is highly unlikely and therefore transformation was done, capping maximum number of hospitalizations at 10.

is transformed variable – original variable capped at 10. In order to achieve linearity in the logit model, variable was further recoded into a three-level variable: 0 – no past hospitalization, 1 – (1-3) past hospitalizations and 2 – (>3) past hospitalizations.

Figure 14. Mortality by number of previous admissions in preceding 12 months



4.3.1.10. Surgery

Dummy variable for patients that underwent surgery during their hospital episode was also created. Surgical procedure was performed in 26.62% hospital episodes, however, mortality in this group is only 1.43%, which is somewhat surprising – I would expect higher mortality as surgery involved significant additional risks to the patient. This low mortality is at least partially explained by the fact that 41.34% of surgeries are classified in the low-risk diagnosis groups (primarily deliveries and other maternal care). Mortality in surgical patients with low-risk diagnosis only is 0.16% and mortality in the remaining surgical patients is 3.33%.

4.3.1.11. Peer groups

Literature on hospital profiling recommends comparing similar hospitals [CIHI, 2012] – major university hospitals are likely to be very different from small regional community hospitals in structure of treated cases. Therefore, despite the fact that I already restricted my analysis to acute care hospitals only, I will consider introduction of a peer group variable in order to account for differences in hospital types.

There are two options of how to approach incorporation of hospital peer group factor into the analysis: introduction of peer group dummy into regression equation or calculation of coefficients for each peer group since in order to perform comparison within peer groups we need to compute predicted mortality on the basis of peer average rather than overall average. First, however, I need to explore whether the peer group factor is a relevant predictor of differences in mortality.

Most obvious solution is to identify two peer groups: university hospitals (11 hospitals that treat 41% of all episodes) versus general hospitals (60 hospitals and 59% of episodes). However, looking at the numbers of hospitalizations, the smallest university hospital has 3 900 cases as compared to 13 750 cases in the largest general hospitals, hinting at large heterogeneity in these groups. Therefore I decided to run a simple cluster analysis[18] in order to identify peer groups of hospitals that are similar in terms of: number of episodes, number of cases transferred to and from the hospital, Charlson comorbidity score and number of spells with high risk diagnosis. I excluded variable of deaths on purpose in order to prevent clustering of low-quality hospitals, the aim was to cluster similar hospitals in terms of structure of treated cases.

Mortality in various peer groups does not differ much with the exception of a slightly higher mortality in the smallest hospitals (LH_Peer=1). Chi-square test is significant, confirming existence of differences in mortality between peer groups (Figure 15), however, when we run univariate logit with only LH_Peer variable, coefficients are not significant (Figure 16).

Figure 15. Cross-tabulation of mortality by peer-groups

LC_DeathIN	LH_Peer 0	1	2	3	4	Total
0	147 98.66	30,919 97.14	84,808 97.58	97,971 97.41	165,970 97.56	379,815 97.49
1	2 1.34	909 2.86	2,100 2.42	2,605 2.59	4,146 2.44	9,762 2.51
Total	149 100.00	31,828 100.00	86,908 100.00	100,576 100.00	170,116 100.00	389,577 100.00

Pearson chi2(4) = 25.8529 Pr = 0.000

[18] K-means cluster with Euclidean measure. Data include also cases transferred from other and to hospital.

Figure 16: Logit model with only one explanatory variable - peer group.

P-values are not significant

```
Logistic regression                                     Number of obs   =     389577
                                                        LR chi2(4)      =      25.37
                                                        Prob > chi2     =     0.0000
Log likelihood = -45614.212                             Pseudo R2       =     0.0003
```

LC_DeathIN	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_ILH_Peer_1	2.160711	1.539879	1.08	0.280	.5345258	8.734227
_ILH_Peer_2	1.819871	1.296149	0.84	0.401	.4506068	7.349936
_ILH_Peer_3	1.954197	1.39169	0.94	0.347	.4839289	7.89142
_ILH_Peer_4	1.835937	1.307281	0.85	0.394	.4547355	7.412366

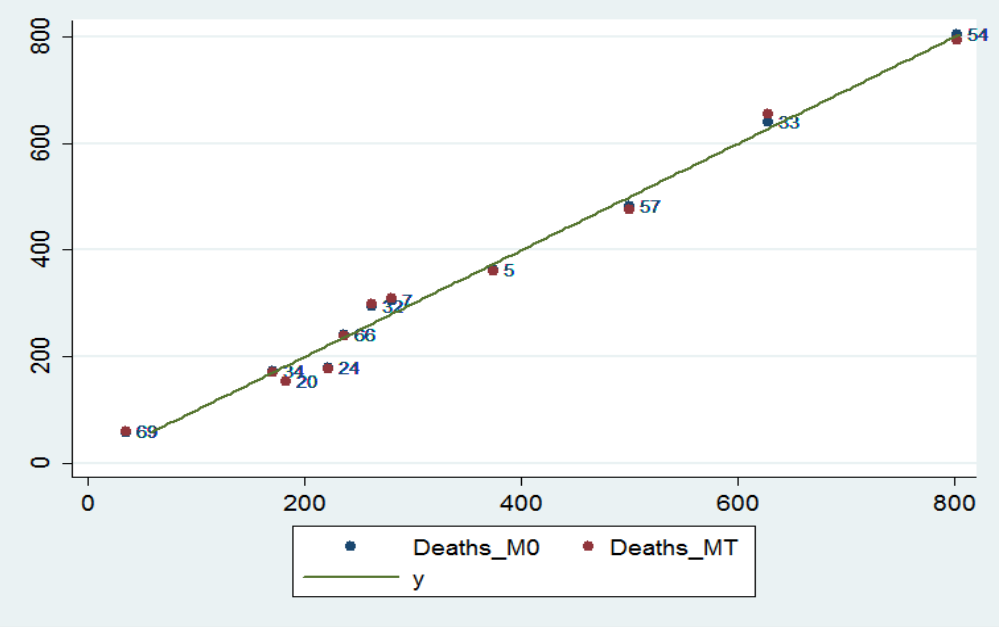
I performed the same analysis for only two peer groups: university versus other general hospitals. While the variable is significant this time and the odds of a patient dying if treated in general hospital are 1.15 times higher than the odds of a patient treated in the university hospital, it is not a particularly useful variable – it increases the measure overall fit of the model - likelihood ratio - by only 44 (for comparison Charlson comorbidity index increases likelihood ratio by 11459.73 and the peer group variable composed on the basis of cluster analysis by only 25). **Therefore, it was decided not to consider peer group factor further on in the analysis.**

Nevertheless, for the illustrative purposes I ran one of the possible risk-adjustment models[19] separately for university hospitals only and compared predicted number of deaths by hospitals against observed number of death with the model run on all the data regardless of the peer type. In Figure 17, we can see the results for university hospitals – differences do exist, particularly for hospital 33 and 54: red dots represent predicted number of deaths for the entire model and blue dots for the model run on university hospitals data only and the green line represents average: observed number of deaths equals to predicted number of deaths. The effect of peer group is not uni-directional – while hospital 33 would decrease its negative differential against average (defined as average in the peer group) should the model for university hospitals be used, hospital 54 would lose its small positive differential. However, given confidence intervals that are not considered at this stage, differences might not be significant at all. Peer groups will not be considered in the following analysis.

[19] logit LC_DeathIN LC_EMERG1_ARO i.ChCI2 DGN_HR DGN_LR LP_Sex LP_Age LC_TranferIN CHA

Figure 17. Predicted number of deaths plotted against real number of deaths by hospitals – university hospitals only.

Blue dots represent risk-adjustment model computed separately for university hospitals, red dots represent model run on data of all the hospitals. Green line represents comparison group average – when predicted number of deaths equals observed number of deaths.



5. Methodology

Providers‘ profiling is usually a two-step endeavour – in the first step, the mortality prediction model is estimated via a binary response model modelling the probability of dying of each treated patient based on relevant risk factors and in the second step the actual mortality is compared against the predicted mortality for every evaluated provider. The resulting index is than interpreted as a departure of the actual mortality from the expected mortality based on reference data – in our case nation-wide mortality in acute care general hospitals. I will employ standard method for derivation of mortality prediction in hospital profiling – logistic regression. Split-sample approach will be applied for validation of the constructed model – model will be estimated on the test sample of 75% of the original data and validated on the remaining 25% - so–called validation sample. During the process of model building, various mortality measures will be compared on the basis of relevant statistical criteria and stability of coefficients across validation sample. In the next step confidence intervals will be calculated and outliers (above and below average hospitals) will be evaluated. It needs to be said that the purpose of this exercise is not to compile an exact ranking from the best to the worst hospital – the best we can hope for with any level of statistical significance is to identify outliers – those hospitals whose performance most likely differs from the average of the reference group. Validation of the overall results of the hospital-profiling exercise (i.e. identification of above average and substandard hospitals) is difficult as no independent evaluation of hospitals is available to serve as benchmark (the only meaningful effort at profiling of a private insurance company is three years old). The only validation I will perform is the comparison of the results from my method against mortality in low-risk diagnosis groups as I consider good performance on this mortality indicator a necessary, even though not sufficient, criterion of hospital quality.

5.1. Logistic regression model

Logistic regression model is a binary response model that allows predicting the probability of a discrete outcome, such as group membership. It is therefore widely used for modelling dichotomous dependent variables, for which it estimates the odds or probabilities of individual cases achieving the outcome. In our exercise, we are modelling binary variable: death or survival of the patient after a hospital episode.

Absolutely crucial task in successful profiling of providers is fitting the mortality prediction model – which variables to include and how should they be transformed and which or whether any interaction variables should be included. There are two options basically – either the variables are selected and tested in the model arbitrarily by the researcher or a stepwise regression is performed automatically by the statistical software – backward stepwise method starts from the full model (i.e. model that includes all the possible explanatory variables), testing deletion of each variable and removing the one which deletion improves the model most or which does not fulfil test for inclusion in the models at the specified significance level (likelihood-ratio test most often)[20]. This is a rather convenient way of doing things, nevertheless there are serious concerns regarding its statistical properties, such as performing multiple comparisons and danger of overfitting the model. Stepwise regression is therefore recommended to be used with extreme caution [Peng and So, 2002].

Given the fact that I did not have that many candidate variables and shortcomings of the automatic techniques, I fitted the model manually. In the first step, existence of differences in mortality for each candidate variable was tested by chi-square test (Figure 6). In the next step we started with the full model and gradually deleted variables that had low addition to the overall goodness of fit and prediction accuracy. We have a number of options of how to evaluate the goodness of fit of the logistic regression model: 1) tests of individual predictors 2) various R-squared measures for overall goodness of fit of the model and again likelihood ratio and 3) various measures of prediction accuracy and corresponding misclassification rates.

P-value for the z-test can be interpreted as a test of significance of individual predictor (same as p-value of t-test in linear regression). However, regarding size of our dataset p-value looses part of its informative value (for details see section 3.6.) as it is significant for most of the variables even though the variable explains very little of the actual variation in the data. Therefore, rather than p-value we will look at l**ikelihood ratio test** which is one of the basic measures for validation of the overall goodness of fit of the model as well as validation of individual predictors and is generally considered superior to other common test for validation of individual predictors – Wald test [Peng and So, 2002]. Likelihood ratio test is based on deviances – the test compares the difference between deviance of the null model without the predictor and model with the predictor on a chi-square distribution given the respective degrees of freedom. Likelihood ratio expresses how many times more likely the data are given the model with the predictor – the higher the ratio the better

[20] Forward selection start the null model and tests for additions of individual variables

the fit of the model with the predictor [Peng and So, 2002]. In addition to evaluating significance of the predictor, likelihood ratio test also gives estimation of the strenght of the predictor's contribution to the model.

Variations of R-squared concept as known from linear regression analysis can be also used for assessing the goodness of fit of the model. However, while R-squared concept has a straightforward interpretation in linear regression – proportion of variance in the dependent variable explained by the model, various pseudo R-squared in logistic regression lack such an easily understandable interpretation. Various authors recommend different measures, usually McFadden's pseudo R-squared is preferred over others [Peng and So, 2002]. It should be also mentioned that R-squared measures tend to lower than what we are used to seeing in linear regression because we are predicting outcomes – either 0 or 1, but what we actually get from the model are only probabilities [Lunt, 2012]. Given the lack of consensus on the R-squared statistics in logistic regression, a set of measures based on predictive accuracy is recommended as a more viable option.

Another type of tests for logistic regression focus on the prediction accuracy of the tested models. Overall **predictive accuracy** expresses the percentage of observations that are correctly classified by the model and partial more specific measures include **sensitivity** (percentage of true positives identified by the model), **specificity** (percentage of true negatives identified by the model), positive predictive value (percentage of true positives out of all positives identified by the model) and **negative predictive value** (percentage of true negatives out of all expected negatives). All of these can be calculated from classification table – one of the primary outputs from the logistic regression analysis. **Hosmer-Lemeshow test** is a test of the overall goodness of fit of the model that compares observed and predicted numbers for various subgroups of data. [Peng and So, 2002] However, application of these measures is somewhat complicated in our case due to the rare events prediction problem and problems associated with large datasets. Prediction accuracy is very high without any explanatory variables because occurrence of the adverse event is extremely rare – mortality is only 2.5% ad therefore 97.5% prediction accuracy is achieved in model-free scenario and H-L test is very sensitive to sample size: in large datasets even small deviations between observed and predicted values yield small p-values [Schwartz et al., 2006]. Furthermore, while these tests evaluate predictive accuracy on the level of individual observations, in the end we are interested in the aggregated predictions – aggregation by the hospitals as these are aggregations of probabilities rather than simple 0-1 values.

While all of the above-mentioned measures asses both calibration and discrimination performance of the model, so called **c-statistics** evaluates only the ability to discriminate between those that experienced the adverse outcome and those that did not [Schwartz et al., 2006]. C-statistics is defined as the area under operator receiving characteristics curve (ROC) that represents the probability that when case of one negative outcome and positive outcome are selected at random, predicted probability for the actual positive outcome is higher than for the negative outcome. Schwartz et el. (2006) argue that even though both calibration and discrimination are crucial properties of a good model, it is relatively easy to re-calibrate the model (cross-validations, different patient populations), but it is harder to improve a model that fails to discriminate properly.

5.2. Preliminary model: selection of variables

Based on the indicators outlined in the previous section, I arrived at a preliminary model that seemed to performed best in terms of the selected indicators: McFadden's R-squared, predictive accuracy, sensitivity, positive predictive value and c-statistics. For comparison of several models see table in Figure 17: Model0 is the null model with no explanatory variables, Model1 includes the very minimum of any risk-adjustment efforts – age and sex Model3 includes 4 variables with the highest likelihood ratios based on the univariate tests and finally Model3a is the preliminary model with 9 variables plus 5 dummy variables for Charlson comorbidity index. It should be noted that this model includes only three diagnosis variables – high risk, medium risk and low risk – which is a rather high level of aggregation. However, model3b which includes 17 separate diagnosis categories interestingly enough performs worse in most criteria than Model3a, primarily in prediction accuracy criteria. Therefore, Model3a is selected as preliminary model.

Figure 18. Comparison of goodness of fit measures for several alternative models

	MODEL0 no explanatory variables	MODEL1 AGE, Sex	MODEL2 AGE, High-Risk, Low-Risk, Charlson Index	**MODEL3a** Age, Sex, Charlson comorbidity Index, TransferIN, emergency transport, emergency ARO, No of previous hospitalizations, **3 diagnosis groups - high risk, medium risk and low risk**	MODEL3b Age, Sex, Charlson comorbidity Index, TransferIN, emergency transport, emergency ARO, No of previous hospitalizations, **17 diagnosis groups**
pseudo R-squared (McFadden's)	0	0.1564	0.2235	0.2917	0.2944
LR chi2(0)	0	10615.66	15169.23	19799.23	19980.83
Prob > chi2	na	0.000	0.000	0.000	0.000
C-statistics	0.5	0.8238	0.8768	0.9027	0.9069
Predictive accuracy	97.51%	97.51%	97.51%	97.57%	
Sensitivity	0%	0%	0.04%	8.29%	8.06%
Specificity	100%	100%	100.00%	99.85%	99.85%
Positive predictive value	0%	0%	75.00%	59.33%	57.92%
Negative predictive value	97.51%	97.51%	97.51%	97.71%	97.56%
Hosmer-Lemeshow test - **chi-square**	na	170.77	109.33	177.64	163.78
Hosmer-Lemeshow test - **p-value**	na	0.000	0.000	0.000	0.000

5.3. Preliminary model: diagnostics

5.3.1. Interaction terms

First of all, we need to consider inclusion of interaction terms – variables that account for differential impact of one variable conditional on another variable (for example different probability of dying for men and women depending on age). Normally, necessity of inclusion of interaction terms can be tested by various tests for specification error that can identify existence of omitted variables or presence of irrelevant variables. However, since the chances that test for specification error in my model would be negative is very low given the fact that I do not risk-adjust for any clinical variables, running specification tests is meaningless.

Considering the variables included in my preliminary model, I suspect that interaction terms might be needed in all the variables in respect to age: probability of dying based on any predictor will be probably different for various ages. However, I will inspect possible interactions only for those variables which are the most important in the model (highest likelihood ratio): interaction between Charlson comorbidity score and age, emergency admission to the unit of intensive care and age and high-risk diagnosis and age. Furthermore, the new variable will be also tested by likelihood ratio test.

Figure 20 depicts impacts of age on the predicted logit (ln(odds)) of dying age for different Charlson comorbidity scores. If no interaction between the two variables exists, the lines on the graph should be paralell because the lines were constructed for regression xi:logit LC_DeathIN i.ChCI2*age that allows for the presence of interaction terms. Differences are obviously significant, particularly for Charlson score of 0 and 5. In case of interaction between age and emergency admission to unit of intensive care within 24 hours, there is no evidence of significant difference – see Figure 21 - LEFT. Figure 21 - RIGHT illustrates that the interaction term between High-risk diagnosis and age is significant – lines are not paralell. For a better idea of the effect of the interaction term, see Figure 22: blue dots represent predicted probabilities of mortality for model including variables age, dummy for High-risk diagnosis and their interaction term, red dots stand for predicted probabilities without the interaction terms and green and yellow dots represent observed mortality by age separately for episodes DGN_HR=1 and DGN_HR=0. Obviously, the blue dots fit the data much better than the red dots: without the interaction term, probability of dying was overestimated for older people and underestimated for people up to 70 years. Differences in the estimated probabilities of dying for a high-risk person with age 0 before and after introduction of interaction term increase from 0.48% to 2.4% and for a person with age 100 estimated probabilities decreased from 48.4% to 28.9%.

Figure 19. Illustration of interaction between Charlson index and age

Ln(odds) of model including interaction term plotted against age. Fact that the lines are not parallel signifies existence of differences in the effect of age for different values of Charlson score

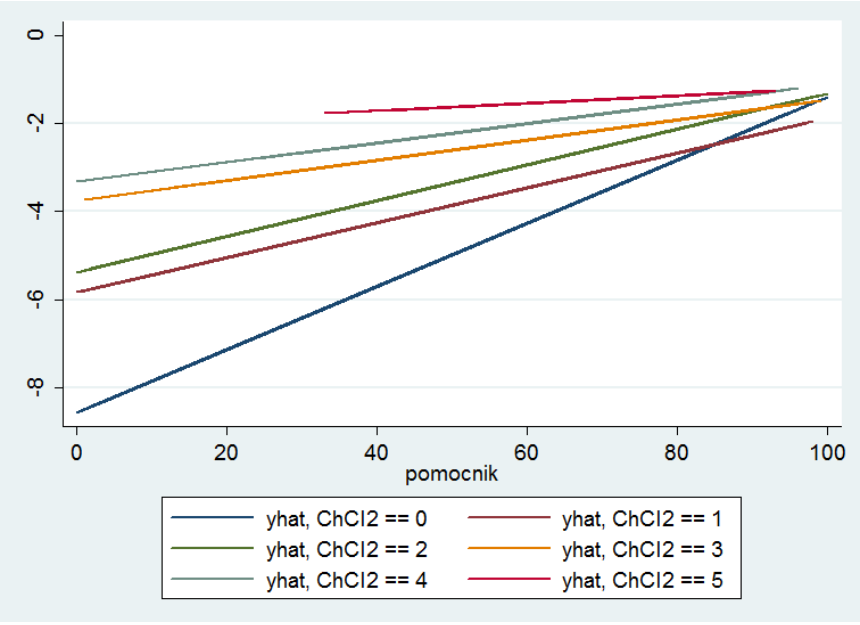


Figure 20. LEFT. Interaction between LC_EMERG1_ARO and age: ln(odds) of model including interaction term plotted against age. Not significant, lines are nearly paralell.

RIGHT. Interaction between DGN_HR and age: ln(odds) of model including interaction term plotted against age. Lines are clearly NOT paralell, effect is significant.

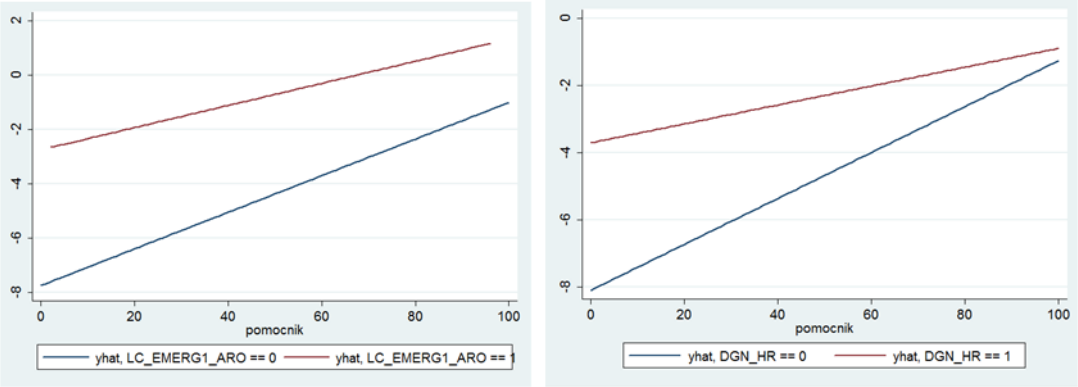
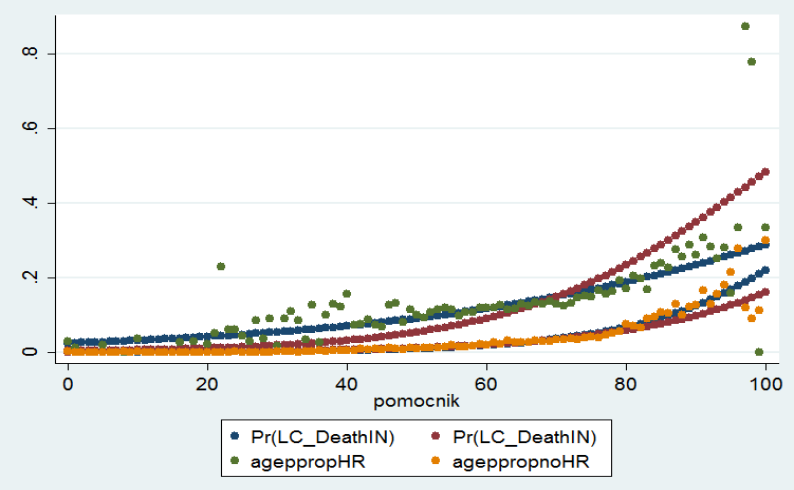


Figure 21. Comparison of predicted probabilities of dying with and without interaction term between age and high-risk diagnosis.

Model of in-hospital mortality with only two explanatory variables: age and high-risk diagnosis. Blue dots represent model with the interaction variable and red dots stand for model without interaction term. Green and yellow dots represent observed mortality by age for episodes for the high-risk diagnosis (green) and without (yellow).



Two new interaction variables will be therefore added to our preliminary model. All the observed measures of the goodness of fit of the model improved by the addition of the two new variables (see Appendix 7).

5.3.2. Linearity

Box-Tidwell regression is test that can check linearity of continuous variables and suggest appropriate transformations. In logistic regression model, relationship between independent and dependent variables are not expected to be linear, but due to the form of link function in the logit model, log(odds) of the variable need to be linear. Figure 23 indicates that three variables are problematic with significant p-values: age, interaction term between age (at this stage age variable was still continuous) and Charlson index (CHA) and number of hospitalizations. Obviously, even though I attempted to transform the age variable to best fit the data, non-linearity is still present. Therefore, I will recode age into 10 dummy variables – 10-year age bands. Variable CHA was transformed as suggested by the test results and square-root was applied and variable LP_NoHosp was recoded from 10 categories to only 3: 0 – no previous hospitalization, 1 - 1-3 hospitalizations, 2 – more than three hospitalizations. Afterwards, only two continuous variables remained in the model – CHA and CHA2 and the Box-Tidwell test failed to reject the null hypothesis of no deviation from linearity.

Figure 22. Box-Tidwell regression

Test for linearity of variables. If variable is linear, p1 equals value of 1

```
-------------+------------------------------------------------------------------
pomocnik |   .0611796   .0014044     43.563   Nonlin. dev. 15.776  (P = 0.000)
      p1 |   1.245572   .0759526     16.399
-------------------------------------------------------------------------------
CHA      |  -.1073002   .0100463    -10.681   Nonlin. dev. 13.454  (P = 0.000)
      p1 |   .4957049   .1221155      4.059
-------------------------------------------------------------------------------
CHA2     |  -.1798132   .0238523     -7.539   Nonlin. dev. 2.390   (P = 0.122)
      p1 |   2.516376   1.301642      1.933
-------------------------------------------------------------------------------
LP_NoHosp|   .0833347   .0062781     13.274   Nonlin. dev. 14.969  (P = 0.000)
      p1 |   .2598301   .2114116      1.229
-------------------------------------------------------------------------------
Deviance:47768.345.
```

5.3.3. Collinearity diagnostics

Standard statistical models can usually tolerate certain amount of multicollinearity between independent variables, but too much inter-dependence in the data might cause inflated standard errors, leading to unreliable estimates of the coefficients in the logistic regression model [IDRE, 2012]. Best practice recommendations is to allow for maximum VIF (variance inflation factor) of 10. Figure 24 below demonstrates that problem exists in 4 variables – collinearity that was introduced into the model by the addition of the two interaction terms. Common solution to this problem is to

center the variable that was used to create the interaction terms – in our case the age variable [IDRE, 2012]. Mean of the variable is deducted from the variable itself and interaction term are than created based on this centered variable. Figure 25 demonstrates that transformation (centering of age) helped to resolve problem of multicollinearity – maximum VIF after transformation is 4.27.

Figure 23 . Collinearity diagnostics before centering

Collinearity Diagnostics

Variable	VIF	SQRT VIF	Tolerance
LC_DeathIN	1.13	1.06	0.8841
LP_Sex	1.05	1.03	0.9501
agegr	1.91	1.38	0.5224
ChCI2	14.75	3.84	0.0678
CHA2p	17.16	4.14	0.0583
CHA2	11.78	3.43	0.0849
LC_TranferIN	1.02	1.01	0.9779
LC_EMERG2_Transport	1.05	1.02	
LP_NoHosp2	1.12	1.06	0.8938
DGN_MR	1.14	1.07	0.8750
DGN_HR	11.52	3.39	0.0868
DGN_LR	1.25	1.12	0.8016
LC_EMERG1_ARO	1.08	1.04	0.9295
Mean VIF	5.07		

Figure 24. Collinearity diagnostics after centering

Collinearity Diagnostics

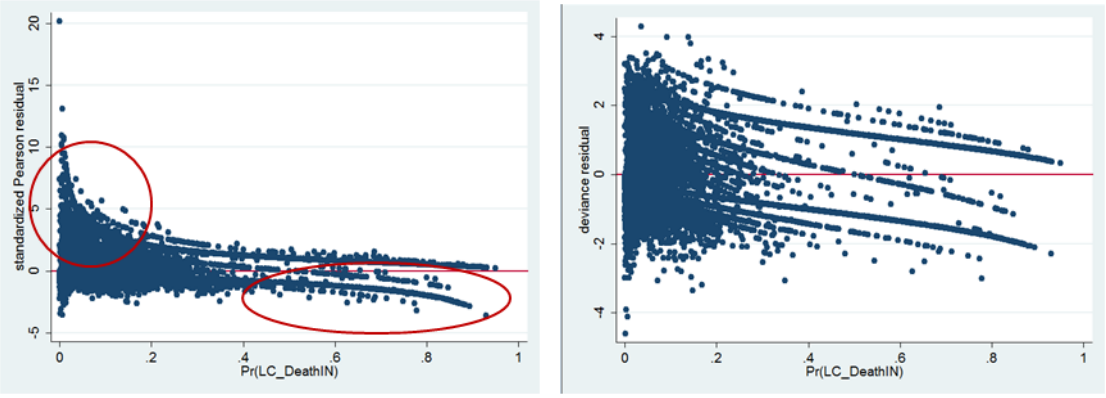
Variable	VIF	SQRT VIF	Tolerance
LC_DeathIN	1.13	1.07	0.8812
LP_Sex	1.05	1.03	0.9484
agegr	1.68	1.30	0.5962
ChCI2	3.49	1.87	0.2865
LC_TranferIN	1.02	1.01	0.9784
CHAc	3.72	1.93	0.2687
CHA2c	2.52	1.59	0.3962
LC_EMERG2_Transport	1.05	1.02	
LP_NoHosp2	1.12	1.06	0.8926
DGN_MR	1.14	1.07	0.8802
DGN_HR	2.49	1.58	0.4019
DGN_LR	1.24	1.12	0.8036
LC_EMERG1_ARO	1.08	1.04	0.9291
Mean VIF	1.75		

5.3.4. Residuals

Examination of residuals might be informative in helping us to learn in which areas the model performs poorly. I will look at two types of residuals – Pearson standardized residuals and deviance residuals. Pearson standardized residuals measure the relative difference between observed and predicted values. Deviance residuals work on the maximum likelihood principle – they compare deviance between maximum likelihood function fitted to the observed and the fitted values [IDRE, 2012]. In Figure 26 - LEFT we might identify two areas where the residuals are somewhat concentrated: Pearson residuals suggest that the predicted probabilities are underestimated (residuals are concentrated on the plus side of the 0 line) in episodes with low prediction scores and a little overestimated for high predictions (residuals are concentrated on the minus side). However, deviance residuals in the Figure 26 - RIGHT do not confirm this hypothesis, as these are distributed rather regularly on the plus and minus side. Therefore, I will not continue further analysis.

Figure 25. LEFT Standardized Pearson residuals plotted against predicted probabilities

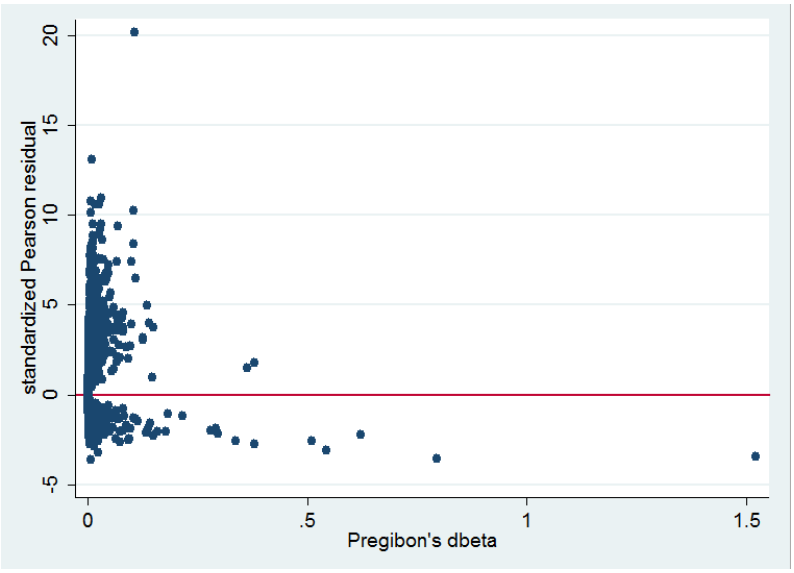
RIGHT Plot of deviance residual against predicted probabilities



5.3.5. Influential observations

Other than errors in the process of the model building, we also need to check for potentially influential observations. First, these might indicate problems in data entry, second we might want to study these observations in more detail to see whether we can detect any pattern and third, influential observation might potentially skew estimates from the logit regression – particularly if there are relatively few observations in certain categories [IDRE, 2012]. Figure 27 depicts standardized Pearson residuals plotted against leverage measured by Pregibon's dbeta – quantity that measures the amount of change in the parameters of logistic regression as the i^{th} observation is omitted [Lunt, 2012]. We can see that while there are some influential observations in the model (high dbeta), these observations do not have particularly high residuals – even though these points have high leverage, their exclusion from the model is unlikely to have significant impact as they are not outliers. Therefore, we do not need to investigate any further.

Figure 26. Standardized Pearson residuals plotted against leverage – Pregibon's dbeta.



5.4. Preliminary model: evaluation

5.4.1. Calibration

In this section, I will evaluate predictive accuracy of the final model we estimated. Crucial tool for analyzing prediction performance of the model is classification table. It summarized the ability of the model to correctly assign group membership to individual observation – by default an observation with predicted probability over 0.5 is classified as 1 – positive group membership. classification table for the model is listed in Figure 28. Overall predictive accuracy is very high – 97.58% which is not surprising given that death is a rare outcome. Model without any explanatory variables is accurate in 97.51% cases because these are all non-events – automatically classified correctly. Sensitivity refers to the percentage of true positives classified correctly by the model. In mour model, we have been able to correctly predict only 8.59% of deaths (using the cutoff point for assigning adverse outcome of 0.5). We must therefore consider the possibility that the estimates are biased – probability of dying is underestimated due to the rare events modelling problem.

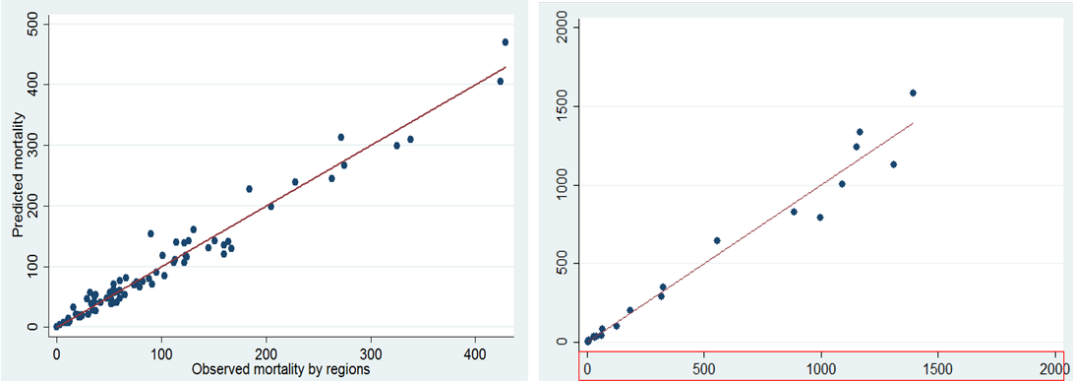
Figure 27. Classification table of the final model

```
Logistic model for LC_DeathIN

                    ———————— True ————————
Classified |              D              ~D |          Total
———————————+———————————————————————————————+———————————————
     +     |            623             414 |           1037
     -     |           6631          283167 |         289798
———————————+———————————————————————————————+———————————————
   Total   |           7254          283581 |         290835

Classified + if predicted Pr(D) >= .5
True D defined as LC_DeathIN != 0
——————————————————————————————————————————————————————————
Sensitivity                     Pr( +| D)    8.59%
Specificity                     Pr( -|~D)   99.85%
Positive predictive value       Pr( D| +)   60.08%
Negative predictive value       Pr(~D| -)   97.71%
——————————————————————————————————————————————————————————
False + rate for true ~D        Pr( +|~D)    0.15%
False - rate for true D         Pr( -| D)   91.41%
False + rate for classified +   Pr(~D| +)   39.92%
False - rate for classified -   Pr( D| -)    2.29%
——————————————————————————————————————————————————————————
Correctly classified                        97.58%
——————————————————————————————————————————————————————————
```

We can test for the presence of this bias by testing for correlation between observed mortality and predicted mortality: if the model compression is present, predicted mortality should be increasing more slowly than observed mortality. D'Hoore et al. (1996) recommends testing for correlation between expected adverse outcomes rate and difference between predicted and observed mortality by hospitals. However, these two variables are not likely to be independent if we assume differences in quality of care provided in individual hospitals exist – low-quality hospital might have high mortality rate and high difference between expected and predicted number of deaths and correlation would therefore be significant (imagine that all bigger hospitals – university hospitals are low-quality, then their high mortality rates given their size will be correlated with high differences in predicted and observed number of deaths). Positive correlation than would not signify problems in the model, but differences in hospital quality. Therefore, I will look at the correlation between observed and expected mortality rate between regions (again, correlation, if detected, might be caused by differences in hospital quality if low/high-quality hospitals are clustered in certain regions, but this effect is likely to be much weaker than in hospitals). For further validation, I will also look at the scatter plot of predicted and observed number of deaths for age groups. If I find evidence that higher observed mortality is correlated with lower predicted numbers, we have problem of a biased model.

Figure 28 LEFT Correlation between observed and predicted number of deaths by regions.

RIGHT Correlation between observed and predicted number of deaths by 5-year age groups.

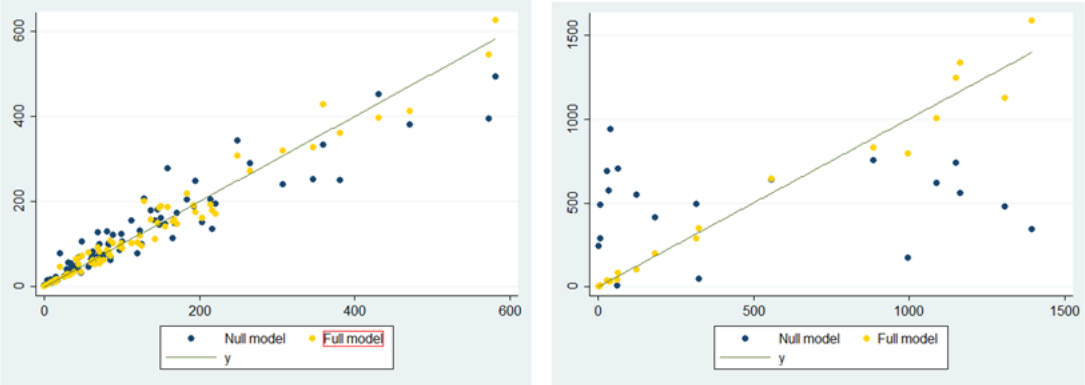


In Figure 29 – LEFT we can see that the relation between predicted and observed mortality is strictly linear – only 4 regions out of 75 have significantly higher predicted probability than observed mortality. Similarly, graph on the RIGHT rejects hypothesis that our prediction model underestimates mortality. Therefore, we can conclude that while it is true that the model predicts only a small proportion of deaths, it is not biased.

Furthermore, this simple analysis suggests that while at the individual level, model predicts only a small number of actual deaths, at the aggregated level – predicting mortality in regions or age groups, the model performs rather well: correlation between observed and predicted deaths is rather high. For comparison of how the mortality prediction model improved predictive abilities at the aggregated level compared to the null model (model with no risk- adjustment) see Figure 30. The graphs plot predicted number of deaths from the full model (yellow dots) and the null model (blue dots) against observed number of deaths. Precision improved significantly. Therefore, there is a good chance that the model should be able to predict overall expected mortality at the hospital level even as it is not recommended to use the model for individual-level predictions – for example for the purposes of identification of high-risk individuals. In this respect, discrimination of the model might be even more important than calibration – indeed, most of the papers dealing with hospital profiling focus on the c-statistics.

Figure 29 Comparison of predicted against observed number of deaths for null model (blue dots – no risk-adjustment) and full model (yellow dots).

LEFT – aggregation by REGIONS, RIGHT – aggregation by AGE groups



5.4.2. Discrimination

As already mentioned, so-called c-statistics is defined as the area under receiver operator characteristics curve (ROC) which is plotted as sensitivity against 1-specificity for every single point. Consequently, maximum value of [0,1] describes situation of perfect discrimination ability of the model – full sensitivity and specificity - value of c-statistics in this instance is 1. C-statistics is interpreted area under the curve - probability that when case of one negative outcome and positive outcome are selected at random, predicted probability for the actual positive outcome is higher than for the negative outcome. In other words, c-statistics captures model's ability to discriminate between events and non-events. Minimum value of c-statistics is 0.5 and it describes a situation when is completely unable to discriminate. Value of 1 represents perfect discrimination. Models with c-statistics between 0.70-0.80 are considered acceptable, c-statistics over 0.80 are considered good [Kansagara et al., 2011]. Figure 31 depicts ROC curve of my final model and it shows exceptionally good discriminative ability of our model with **c-statistics of 0.9071**.

Figure 30 ROC curve for the final model. C-statistics of 0.9071

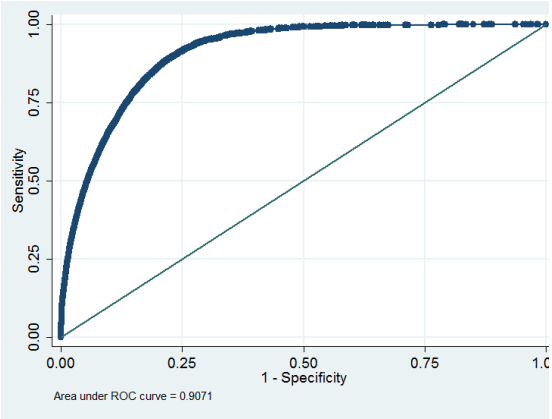
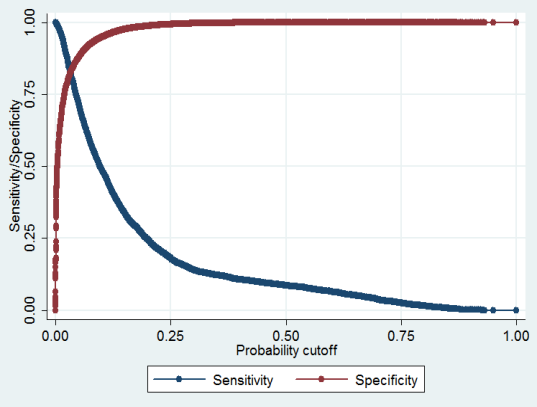


Figure 31. Sensitivity versus specificity plot – optimal cutoff point at 0.033.



Another type of plot often used in the post-estimation analysis of logistic regression models is sensitivity/specificity used for the identification of optimal cutoff point that maximizes both statistics sensitivity and specificity. In my model (Figure 32), the optimum cutoff point for identification of positive cases is extremely low – 0.033 rather than the standard of 0.5 (this value describes the required probability for assigning the case as positive – usually 50% probability is required to identify the case as event). At this point both specificity and sensitivity are at 82.33%, however, the overall predictive accuracy dropped down significantly as we had to decrease specificity in order to increase sensitivity - in other words we had to trade our ability to identify non-events correctly, which was automatic given the rarity of death, for the ability to identify events correctly.

5.5. Validation of the model

Validation of a model is a crucial step in model building, particularly in case of predictive modelling in order to prevent problem of overfitting – situation when predictions do not generalize well to other similar data as they were tailored to fit one particular dataset [Schwartz, 2006]. Validation refers to the process of assessing how well the results generalize to independent data. In case of mortality prediction model, I need to see that coefficients (odds) calculated on dataset of historical hospitalizations are fairly stable so that they can be used to predict mortality on subsamples of data.

In this analysis, I used 75:25 split sample approach – 75% of eligible data were randomly assigned to the derivation sample, on which I performed the analysis and

remaining 25% were assigned to the validation sample. A step further is cross-validation method, whereas the analysis is performed repeatedly on random subsamples of data and the validation results are then averaged over the individual rounds. This approach is strictly non-parametric, it relies on the raw computing power – you repeat the analysis as many times as possible and see distribution of the results of interest. We might consider using a stratified approach to validation – making sure that validation and derivation sample have similar composition in relevant risk factors – in Figure 33 we can see that when we draw completely random samples, composition might differ. However, since chi-square statistics tested negative for statistically significant differences in the composition of derivation and validation sample stratified approach was not deemed necessary.

Figure 32. Comparison of prevalence of individual variables in derivation (Sample_Est==1) and validation data (Sample_Est==0)

Summary statistics: mean
by categories of: Sample_Est

Sample_Est	LC_Dea~N	LP_Sex	agegr	ChCI2	LC_Tra~N	LC_EME~t	LP_NoH~2	DGN_MR	DGN_HR	DGN_LR
0	.0254011	1.581551	3.873187	.5985355	.0236489	.196362	.3428233	.0761424	.0828674	.3869207
1	.0249415	1.583037	3.860711	.5952496	.0232395	.1967432	.3405916	.0753229	.0815222	.3868712
Total	.0250579	1.58266	3.863873	.5960824	.0233433	.1966466	.3411572	.0755306	.0818631	.3868837

Sample_Est	LC_EME~O
0	.0083354
1	.0086405
Total	.0085631

There are two closely related ways of looking at the validation of the results using the split sample approach. First, I will look at how well the estimates obtained from the derivation dataset perform on independent data - is the prediction accuracy maintained? This approach mirrors the way the model will be used in the hospital profiling: coefficients for risk factors will be calculated on the entire dataset, but than predicted probabilities will be applied to sub-samples of data – defined as sum of episodes for each hospital hospitals. If we are unable to conclude that estimated coefficients perform reasonably well on sub-samples of data (in case of validation we are looking at independent but related dataset), we have a problem.

In the second part, I will be interested in the stability of the coefficients across the two samples – I will calculate the coefficients separately on the validation data and compare the coefficients and primarily the predictions. Both steps are two sides of the same coin of course – if the coefficients differ significantly, predictive accuracy when coefficients from one sample are applied to the other one will be low. If the coefficient are not sufficiently stable, I either selected unsuitable model (problem called „overfitting“ – the model predicts mortality well in the training data, but the

variables themselves might not be a good predictors of mortality in general) or high mortality variance of individual episodes, particularly in low-number categories, might cause fluctations of the coefficients. If the latter is the case, coefficients should be routinely calculated by cross-validation and averaged over the individual rounds in order to ensure best results. If the former is the case, cross-validation should be used for the selection of the model itself – model with the most stable results in terms of goodness of fit across various cross-validation samples should be selected.

Figure 34 compares measures of goodness of fit for the derivation and validation sample. We can see that the coefficients applied to the validation sample perform somewhat worse, but discriminative ability expressed by c-statistics remained very high, sensitivity declined a little, however specificity remained unchained and positive predictive value increased. Overall, I conclude that the model performs reasonably well on independent dataset.

Figure 33. Validation of the model: comparison of measures of goodness of fit for the derivation and validation sample

	DERIVATION SAMPLE	VALIDATION SAMPLE
pseudo R-squared (McFadden's)	0.2939	0.2894
LR chi2(0) *	na	na
Prob > chi2	0.000	0.000
C-statistics	0.9071	0.9057
Predictive accuracy	97.58%	97.53%
Sensitivity	8.59%	7.54%
Specificity	99.85%	99.87%
Positive predictive value	60.08%	60.58%
Negative predictive value	97.71%	97.64%
Hosmer-Lemeshow test - **chi-square***	na	na
Hosmer-Lemeshow test - **p-value**	0.000	0.000

**not applicable when comparing model with different number of observations*

Moving on to compare stability of predictions obtained from models calculated separately on derivation and validation sample, Figure 35 - LEFT compares predicted from the validation sample data against probabilities calculated from the derivation dataset (for comparison of calculated odds see Appendix 8). Particularly rworrisome is the bias for high probabilities (>0.6) – predicted probabilities form validation data are higher than predicted probabilities from the derivation data. In other words, the same person has higher odds of dying if we use results from validation data – mean predicted probability on the basis of derivation data is .024982 and for validation data .02527 (difference of 1.1%). However, prevalence of episodes

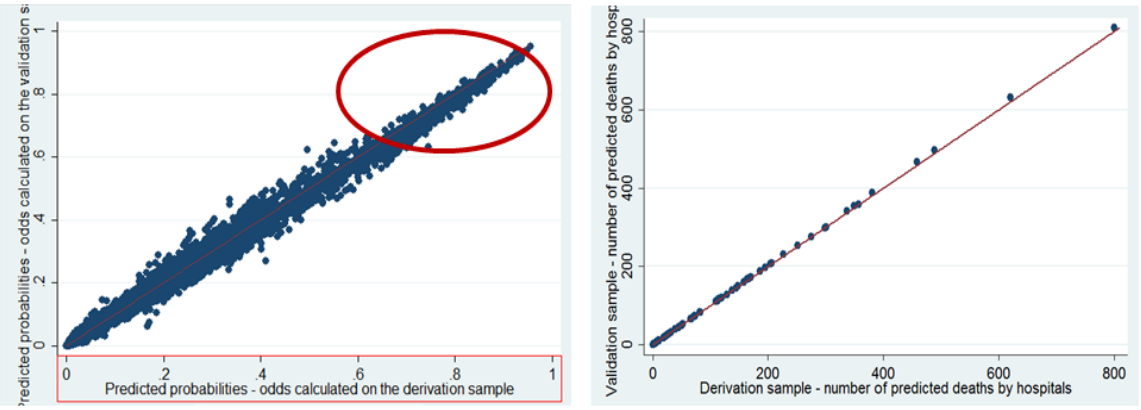
with such a high probability of dying is very low – only 999 individuals out of 389,577. Another way to look at the results is to compare predicted number of deaths for individual hospitals – while predicted probabilities for individual cases might differ (I already pointed out that the model is probably not suitable for prediction at the individual level), results at the aggregated are not quite so disturbing – see Figure 35 - RIGHT. Predicted number of deaths calculated from derivation and validation sample for hospitals differ very little. Some differences exist in case of large hospitals with high numbers of predicted deaths, but the percentage difference is negligible due to high numbers of deaths.

Therefore, conclusion is that cross-validation for the derivation of coefficients is desirable – coefficients are not stable in the derivation and validation sample, but is not absolutely necessary for the purposes of hospital profiling, particularly if we use confidence intervals for the estimation of hospital outliers.

Figure 34 Stability of predictions across derivation and validation samples.

LEFT - scatter plot of predicted probabilities calculated on the basis of validation sample against predicted probabilities calculated on the basis of derivation dataset.

RIGHT Comparison of predicted number of deaths calculated from the validation sample against predicted number of deaths calculated from the derivation sample for each hospital.



5.6. Comparison of various mortality measures

So far I focused exclusively on in-hospital hospital-wide mortality. In this section, I will briefly consider other possible measures of mortality. I examine performance of each mortality measure in terms of model developed for in-hospital hospital-wide mortality. If a variable is found insignificant, it is excluded from the model, however, I did not attempt to fit the best combination of explanatory variables for each

mortality measure – even though that is the ideal option. Comparison of the mortality measures was performed on the basis of the accuracy prediction measures and stability of the model across derivation and validation sample – we saw that in-hospital mortality performed rather poorly in the latter and in order to confidently implement this measure, cross-validation is recommended. Results for prediction accuracy are summarized in Figure 36 and stability is examined in Figures 37 – 42. Not surprisingly, there is no single measure that would perform consistently well in all the specified criteria. However, it can be seen that hospital-wide measures of mortality perform better than more specific measures such as high-risk mortality (comparatively poor c-statistics of 0.7229), mortality for planned hospitalizations only (where overall prediction accuracy actually decreased). Mortality after surgery performed relatively well, but was unstable across validation sample. Therefore, for the purposes of hospital profiling I will consider only three hospital-wide mortality measures – in-hospital, 30-day and 90-day mortality.

Figure 35. Comparison of various mortality indicators: prediction accuracy measures.

Green colour indicates best performing and red worst performing measure for the particular criterion

	In-hospital mortality	30-day mortality	90-day mortality	High-Risk conditions (30-day)	Mortality after surgery (30-day)	Planned hospitalization (30-day)
C-statistics	0.9067	0.9014	0.8970	0.7229	0.9364	0.9081
Sensitivity	8.27%	7.57%	9.37%	13.28%	13.15%	3.28%
Specificity	99.86%	99.78%	99.54%	97.86%	99.82%	99.94%
Positive predictive value	59.87%	58.83%	55.04%	60.17%	60.74%	48.44%
Negative predictive value	97.69%	96.35%	94.84%	82.28%	98.21%	98.40%
Correctly classified - overall predictive accuracy	97.56%	96.16%	94.45%	81.32%	98.04%	98.34%
- increase in % from null model	0.17%	0.07%	0.10%	0.68%	0.10%	-0.12%
R-squared	0.2925	0.2978	0.3046	0.1101	0.3653	0.2720
Stability across derivation and validation samples	-	+	++	--	--	---

Figure 36. Stability of 30-day mortality model predictions

LEFT individual predictions for each observation: validation sample plotted against derivation sample.
RIGHT aggregated predictions for hospitals: validation sample plotted against derivation sample

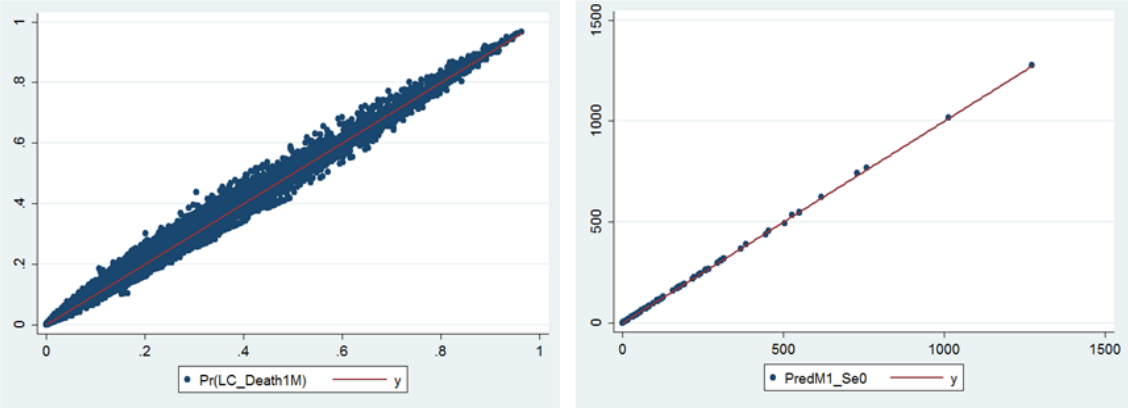


Figure 37. Stability of 90-day mortality model predictions – very stable

LEFT individual predictions for each observation: validation sample plotted against derivation sample.
RIGHT aggregated predictions for hospitals: validation sample plotted against derivation sample

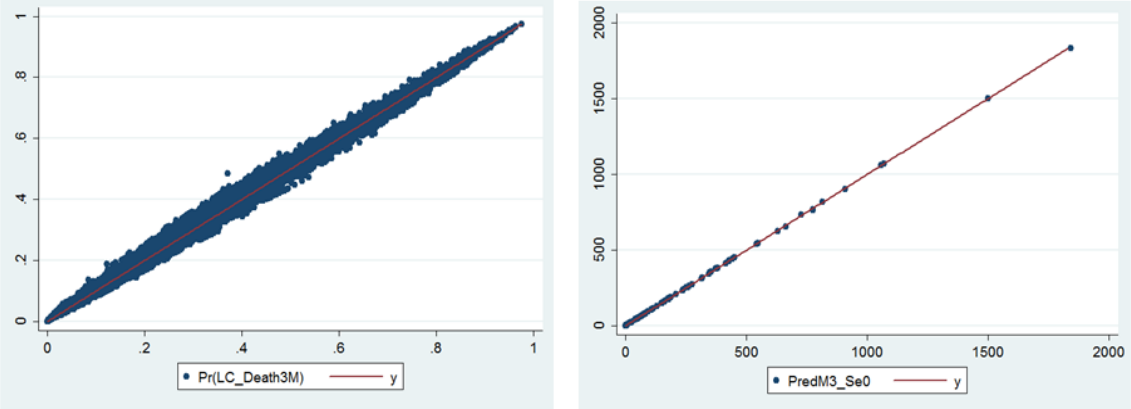


Figure 38. Stability of high-risk mortality model predictions (in-hospital mortality) – wide dispersion

LEFT individual predictions for each observation: validation sample plotted against derivation sample.
RIGHT aggregated predictions for hospitals: validation sample plotted against derivation sample

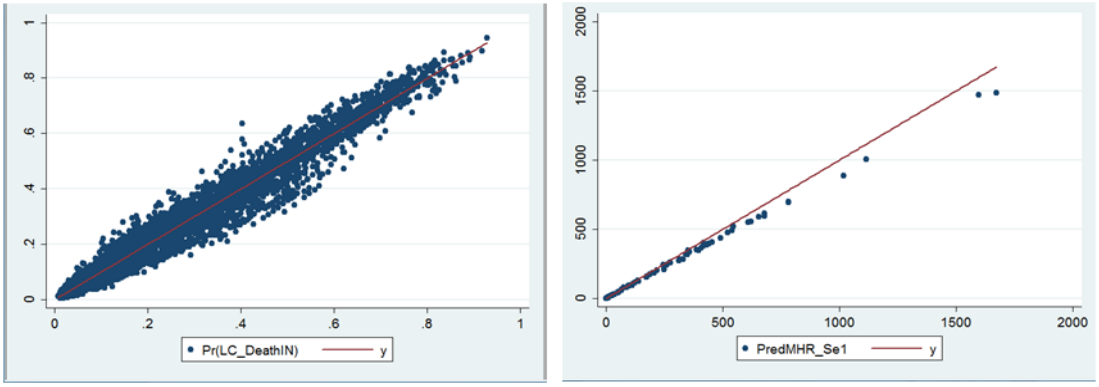


Figure 39. Stability of high-risk mortality model predictions (30-day mortality) - wide dispersion remains

LEFT individual predictions for each observation: validation sample plotted against derivation sample.
RIGHT aggregated predictions for hospitals: validation sample plotted against derivation sample.

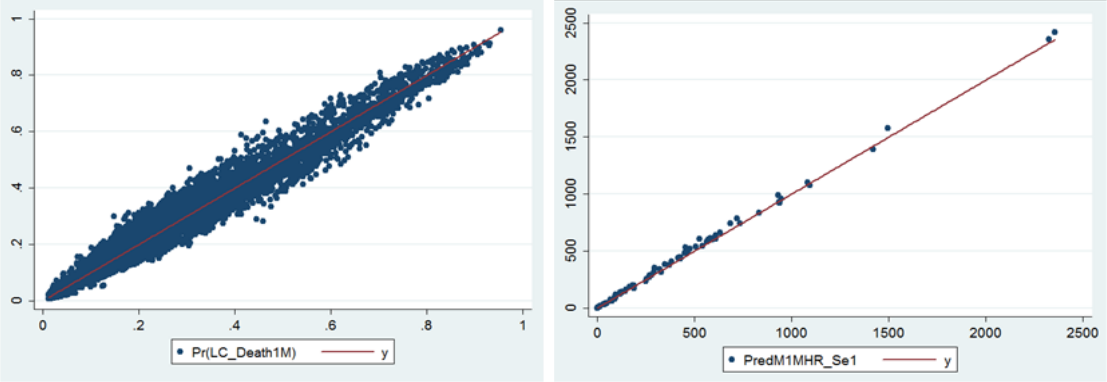


Figure 40. Stability of death after surgery model predictions (30-day mortality) - very wide dispersion and biased

LEFT individual predictions for each observation: validation sample plotted against derivation sample.
RIGHT aggregated predictions for hospitals: validation sample plotted against derivation sample

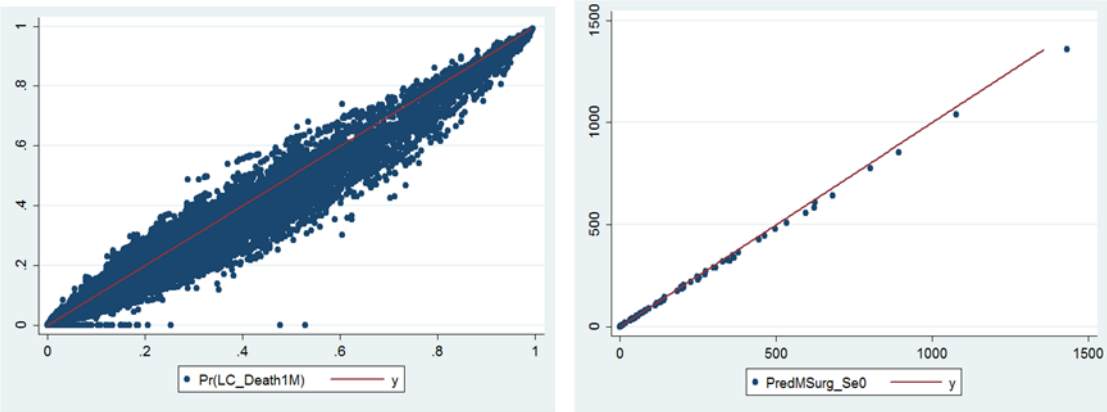
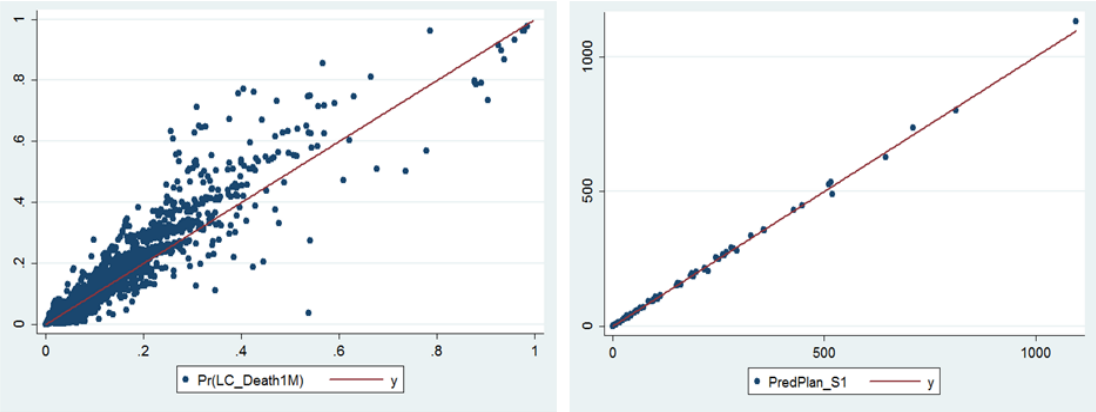


Figure 41. Mortality after planned hospitalizations (30-day) – extremely unstable

LEFT individual predictions for each observation: validation sample plotted against derivation sample. RIGHT aggregated predictions for hospitals: validation sample plotted against derivation sample



5.7. Results – description and interpretation of the final mortality prediction model

Logit model takes the following form [Burns and Burns, 2009]:

$$\boldsymbol{\pi = P(Y = 1|x) = \frac{e^{(\beta_0+\beta_1+\cdots)}}{1+e^{(\beta_0+\beta_1+\cdots)}}} \qquad (1)$$

, where the right-hand side is the logistic cumulative distribution function and β stands for the regression coefficients. Consequently:

$$\boldsymbol{g(x) = logit = \ln\left(\frac{P(Y=1|x)}{1-P(Y=1|x)}\right) = \beta_0 + \beta_1 + \cdots \quad and} \qquad (2)$$

$$\boldsymbol{odds = \frac{P(Y=1|x)}{(1-P(Y=1|x))} = e^{(\beta_0+\beta_1+\cdots)}} \qquad (3)$$

The last equation says that the effect of the predictor on the probability of dying can be expressed by the odds ratios defined as the probability of the outcome given that the predictor is positive divided by the probability of outcome if the predictor is negative: odds=p/(1-p). In other words, in case when the Sex variable increases by 1 (i.e. patient is female), the odds of adverse outcome increase by 0.74 – the odds of dying are lower for women than odds of dying for men. In case of dummy variables for Charlson index, the odds express the increase in the odds of dying compared to the odds of dying of patients with Charlson index of 0.

Figure 43 summarizes the final mortality prediction model calculated on the entire dataset – 389,577 episodes. The model includes several dimensions of risk-adjustment – several categories of variables that are significant factors for mortality prediction: demographic variables, diagnosis groups, comorbidities, past healthcare use and severity of the episodes. One variable is insignificant based on Likelihood ratio chi-square test and other four variables have very low likelihood chi value indicating that they explain little variation in the data. Three of these are age variables – they were left in the model in order to ensure completeness of the data (even though we might consider recoding age 10-39 into one dummy variable) and the other two problematic variables are the interaction terms. However, we demonstrated that these variables have their place in the model in order to ensure better prediction accuracy.

All of the variables increases the odds of the adverse outcome (i.e. odds are higher than 1) with the exception of gender variable, low-risk diagnosis group, age between 10-19 and the interaction terms. The odds of dying are most influenced by age, high-risk diagnosis and indicators of the severity of the episode, particularly admission to the unit of intensive care within 24 hours.

Figure 42. Final risk-adjustment model – in-hospital mortality prediction calculated on the entire dataset.

FINAL MODEL							
Variable	**Description**	**Odds**	**P-value**	**95% CI - lower**	**95% CI - upper**	**LR chi**	**LR test - p-value**
Demographic variables							
LP_Sex	1 - male, 2 - female	0.74	0.000	0.71	0.78	167	0.000
age1	age 10-19	0.33	0.001	0.17	0.65	13	0.000
age2	age 20-29	1.5	0.818	0.71	1.54	0	0.818
age3	age 30-39	2.73	0.000	1.97	3.78	43	0.000
age4	age 40-49	5.44	0.000	4.1	7.38	177	0.000
age5	age 50-59	8.37	0.000	6.22	11.26	370	0.000
age6	age 60-69	12.85	0.000	9.56	17.27	619	0.000
age7	age 70-79	20.39	0.000	15.14	27.46	913	0.000
age8	age 80-89	45.60	0.000	33.73	61.64	1 520	0.000
age9	age >90	87.14	0.000	63.15	120.24	1 257	0.000
Diagnosis							
DGN_HR	high risk diagnosis group	4.41	0.000	3.98	4.89	720	0.000
DGN_MR	medium risk diagnosis group	1.95	0.000	1.84	2.8	420	0.000
DGN_LR	low risk diagnosis group	0.45	0.000	0.41	0.50	277	0.000
Comorbidities							
ChCI1	Charlson score 1-10	1.83	0.000	1.69	1.99	215	0.000
ChCI2	Charlson score 11-20	2.39	0.000	2.14	2.66	242	0.000
ChCI3	Charlson score 21-30	3.73	0.000	3.21	4.33	281	0.000
ChCI4	Charlson score 31-40	5.99	0.000	4.92	7.29	293	0.000
ChCI5	Charlson score >40	9.17	0.000	6.97	12.7	223	0.000
Past use of healthcare services							
LP_No Hosp	Number of past hospitalizations 0 - (0), 1 - (1,3), 2 - (>3)	1.26	0.000	1.22	1.30	200	0.000
Severity of the episode							
LC_TransferIN	admitted transfer	2.35	0.000	2.16	2.56	339	0.000
LC_Emerg2_transfer	emergency transfer to the hospital	2.27	0.000	2.17	2.37	1 231	0.000
LC_Emerg1_ARO	admission to the unit of intensive care within 24 hours of admission to the hospital	12.94	0.000	11.84	14.14	2 884	0.000
Interaction terms							
CHAc	age*Charlson score recoded - centred	0.62	0.000	0.55	0.70	60	0.000
CHA2c	age*DGN_HR -centred	0.86	0.000	0.83	0.89	71	0.000

6. Hospital profiling

6.1. Methodology

The methodological approach to hospital profiling is based on the standardized mortality ratio model whereas observed number of deaths in the hospital is compared to expected numbers of deaths (hospital standardized mortality ratio):

$$HSMR_h = \frac{D_h}{E_h} = \frac{\sum d_i}{\sum p_i} \qquad (4)$$

, where D_h and E_h are the observed and expected number of deaths in the hospital, expected number of deaths is calculated as a sum of probabilities of death (p_i) calculated from the mortality prediction model on the bases of equation (1). and d_i is a binary variable for each observation – 1 is the code for death 0 - survival. The standardized mortality ratio can attain values from 0 to infinity. If the ratio equals to 1, expected and observed numbers of deaths are equal which means that the mortality in the hospital is exactly average in respect to the reference population – in our case reference population being all the episodes in acute-care hospitals in Slovakia. In other words, the ratio captures deviation of mortality observed in the hospital from national average hospital mortality. This interpretation is enabled by the fact that expected mortality is computed on the basis of predicted probabilities from the mortality prediction model that was computed on the data of all acute-care hospitals. These predicted probabilities therefore express the „average". If the HSMR>1, observed mortality in the hospital is higher than expected and we suspect that hospital might be sub-standard.

Crucially important is the calculation of confidence intervals for our estimates - if the calculated HSMR for a particular hospital is 1.1 can we really say that the hospital is better than the average? Confidence interval can be defined as an interval that captures the true value of the parameter of interest at the required confidence level (usually 95%). More precisely, should we repeat the measurement on 100 independent samples, proportion of confidence intervals that includes the true value is 95%. It is this second definition that I will use to construct confidence interval around the point estimate – HSMR as we will see later on.

There are several options of how to calculate confidence intervals including: confidence intervals based on approximation with normal distribution, Byar's approximation, exact Test based on Poisson distribution (these can be used for the construction of control limits rather than confidence intervals and calculations of funel plots) or calculation of confidence intervals using Monte Carlo simulations. Soe and Sullivan (2006) compared various methods for calculation of confidence intervals for standardized mortality ratios and concluded that exact test should be used in cases when expected/observed number of deaths is lower than 5 and for larger numbers, approximations perform comparably well. Below are formulas for two methods of calculating confidence intervals, for more methods see [Soe and Sullivan, 2006].

Confidence interval for dichotomous data for large enough numbers of death can be approximated with normal distribution [Schwartz et al., 2006 and Itskovitch and Roudebush, 2010]:

$$\text{S.E.} = \sqrt[2]{\frac{\sum p*(1-p)}{n}} \qquad (5)$$

$$CI_{lower} = HSMR * \left(1 - \frac{1}{2\sum p*(1-p)} - Z\left(\frac{S.E.}{\sqrt[2]{\sum p*(1-p)}}\right)\right) \qquad (6)$$

, where p – estimated probability of death for each case, n – number of cases treated by the provider, Z – normal distribution score for desired confidence interval (1.96 for 95% CI).

However, as I already struggled with the problem of rare outcomes - overall mortality is only 2.5%, I must look to other alternatives for calculation of confidence intervals, such as Byar's approximation which is a popular choice in hospital profiling models [Itskovitch and Roudebush, 2010]:

$$CI_{lower} = \frac{D}{E} * \sqrt[3]{\left(1 - \frac{1}{9*D} - \frac{Z}{3*\sqrt[2]{D}}\right)} \qquad (7)$$

$$CI_{upper} = \frac{D+1}{E} * \sqrt[3]{\left(1 - \frac{1}{9*(D+1)} + \frac{Z}{3*\sqrt[2]{D+1}}\right)} \qquad (8)$$

However, notice that Byar's approximation does not take into account number of cases treated in the hospital, only the number of deaths.

Finally, there is one option for calculating confidence interval that has a fundamental advantage compared to other methods - it is completely non-parametric, it does not require any assumptions regarding minimum number of events or underlying distributions. Furthermore, conventional approaches to calculation of confidence intervals are based on the assumption that distribution of deaths in the population is completely random and can have any value according to the underlying distribution. This might, however, result in too conservative estimates of confidence intervals that are unjustifiably wide, particularly in the cases that we are interested in a short-time horizon where we know the actual number of deaths [Itskovitch and Roudebush, 2010].

Therefore I will use monte-carlo simulations – randomly drawing subsamples of data and repeatedly calculating HSMRs with the restrictions that the randomly drawn subsamples have the same overall mortality as the real mortality. We might consider using more stratified approach – imposing more restriction on the composition of the random subsamples, which would result in narrower confidence interval, but at this stage I will use just the one restriction. I will compensate for the fact that resulting subsamples might be significantly different in composition of episode characteristics used for risk-adjustment by selecting a relatively large subsamples – 80% of the entire dataset. Nevertheless, this simulation technique should still be able to filter away much of the statistical uncertainty in the data – bad luck that one provider treated a number of cases that are not covered by the risk-adjustment model, for example. Another advantage is that it has an easy intuitive interpretation, possibly better than confidence intervals and its error type 1 and error type 2 explanations. It can help with refuting objections that we cannot say for certain that one hospital is worse than others. This objection is valid – we cannot be certain, but we have strong evidence that this is indeed so: out of 500 risk-adjusted samples of the hospitals' episodes, 95% of these subsamples had above-average mortality.

For the purposes of constructing hospital profiling, I will use data for 12 months – 2011.07 - 2012.06. Predicted probabilities will be estimated based on the in-hospital mortality prediction model developed in the previous section and hospital standardized mortality ration will be calculated (see Appendix 9 for calculated coefficients used in the hospital profiling). This will provide the point estimate around which confidence intervals will be constructed by calculating ratios for randomly drawn 500 subsamples of 80% of the entire dataset (118,000 episodes),

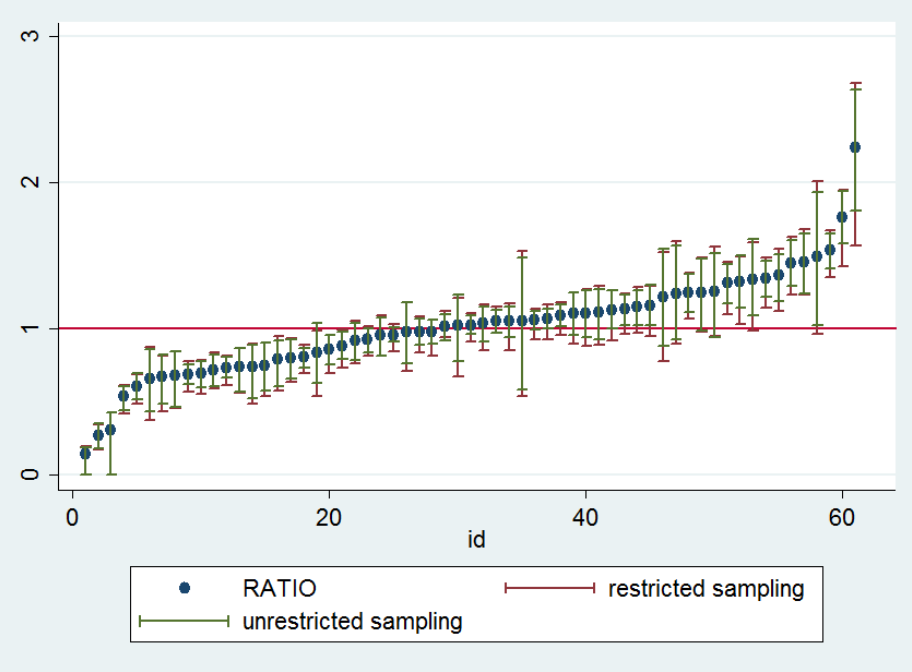
with the restriction that mortality in each subsample is equal to the mortality of the entire dataset (.0259995). Lowest and highest 2.5% of simulations will be discarded in order to obtain 95% confidence interval. This approach to calculation of confidence intervals will be compared with the Byar's approximation technique that is often recommended for the purposes of hospital profiling efforts.

6.2. Results

6.2.1. Comparison of confidence intervals

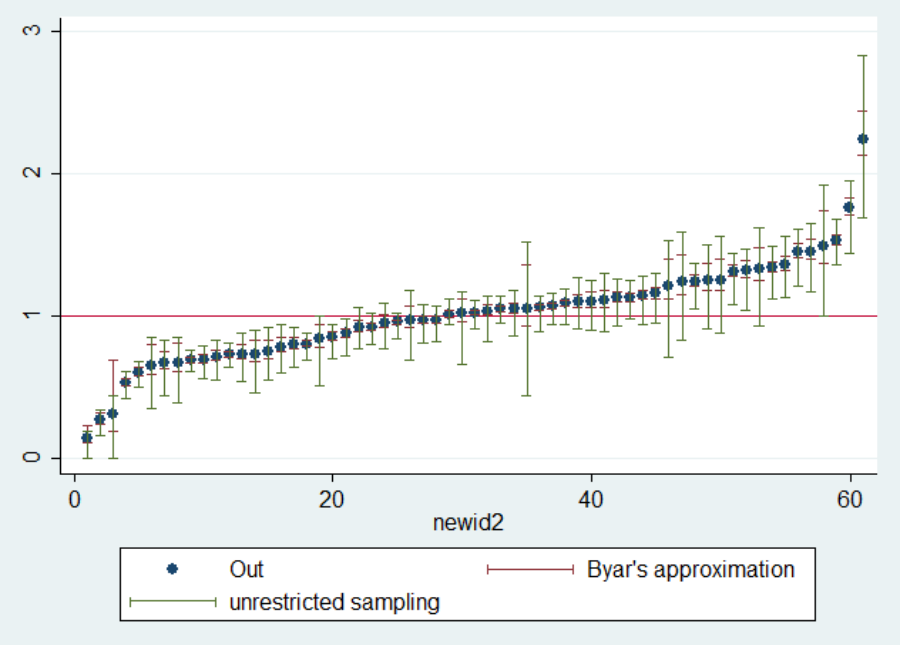
Surprisingly enough, when I used unrestricted sampling (i.e. regardless of the overall mortality in the subsamples), I obtained narrower confidence intervals than when I used restricted sampling – see Figure 44. Originally, I expected to see narrower confidence intervals when I impose additional restrictions on the subsamples – as the variance of values should be lower. However, after closer examination it becomes apparent, than even though the intervals are wider when using restricted sampling technique, they are stricter in identification of outliers – primarily in identification in below-average hospitals. Potentially sub-standard hospitals are those above the horizontal line: their observed mortality is higher than expected mortality. Confidence intervals constructed with the restricted sampling technique are extended downwards – forcing us to conclude that the hospital is not significantly better than average hospital – 5 hospitals were re-classified from below-average to average because of this. All in all, when using restricted sampling for the construction of confidence intervals we receive following results: 10 below-average hospitals (16.4%), 21 above-average hospitals (31.4%) and 30 hospitals that can be considered average (49.2%). In comparison, when using unrestricted sampling I classify 18 hospitals as below-average (29.5%), 20 hospitals as above-average (32,8%) and 23 hospitals that are not significantly different from average (37.7%).

Figure 43. Comparison of confidence intervals calculated with restricted (red lines) and unrestricted (green) sampling



Confidence intervals constructed with Byar's approximation are much narrower and consequently classify hospitals as above-average or substandard much more easily. For graphical comparison of CI calculate using Byar's approximation and restricted sampling see Figure 46 and for the impact of the choice of method on the classification of hospitals see Figure 47.

Figure 44. Comparison of confidence intervals: Byar's approximation vs restricted sampling



This brief analysis of confidence intervals also explains why I will not attempt to construct „hospital ranking“ that ranks hospital from the best one to the worst one. Instead, I will merely attempt to separate good hospitals from the average hospitals and average hospitals from the bad ones. Confidence intervals reflect the measure of uncertainty in our point estimates (i.e. HSMRS), which is substantial. Looking at Figure 46, where we can actually see the ranking from best performer to the worst one, it becomes instantly obvious that after considering confidence intervals neighbouring hospitals become easily interchangeable in their respective ranks.

Figure 45 Impact of confidence interval calculation method on hospital ranking

Impact on confidence interval calculation method on the hospital ranking			
	Byar's approximation	**Unrestricted sampling**	**Restricted sampling**
Below-average	30 (49.2%)	18 (29.5%)	11 (18.0%)
Average	7 (11.5%)	23 (37.7%)	30 (49.2%)
Above-average	24 (39.3%)	20 (32.8%)	20 (32.8%)

**number of hospitals classified in each category (%of hospitals classified in the category)*

6.3. Impact of risk-adjustment on hospital ranking

In this section I will finally look at the question that inspired my research: does risk-adjustment matter? How does the introduction of risk-adjustment influence classification of individual hospitals? What impact does selection of mortality indicator have on the outcome of the profiling?

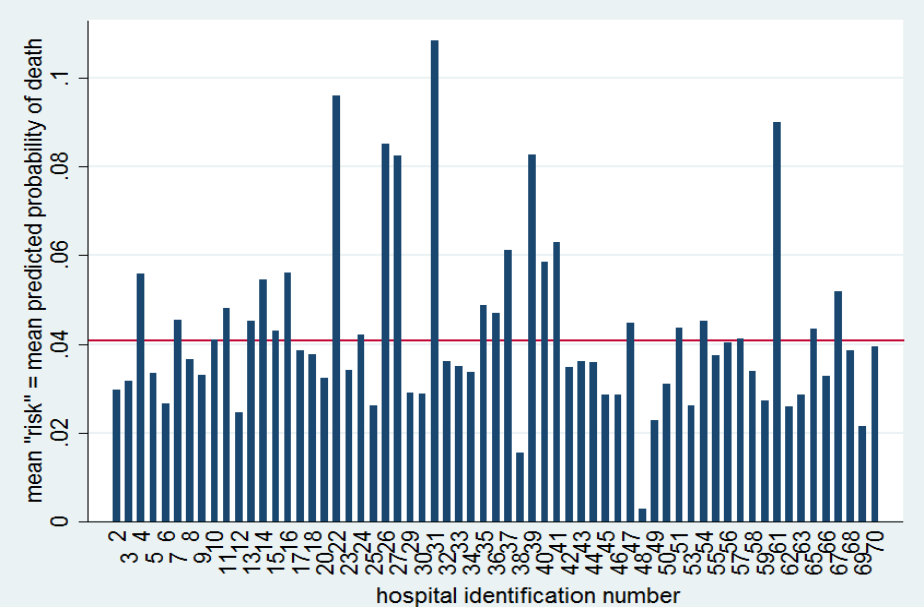
6.3.1. Effect of introducing risk-adjustment on hospital classification

First of all, before we proceed to the actual hospital profiling, let us verify our original assumption that motivated this research in the first place – risk-adjustment is crucial for hospital profiling because different hospitals have different risk structure of treated patients. We developed risk-adjustment model and evaluated its prediction accuracy and validated the model across validation sample, but we also need to look at whether the model distinguishes between hospitals. Figure 47 plots mean predicted probability of death that can be also interpreted as mean risk of the patient population for each hospital. Differences in the “riskiness” of patient populations are obvious and quite pronounced for certain hospitals: mean predicted probability of death

differs from 0.02% to 33%. The graph not only demonstrates that risk-adjustment is needed, but also that we succeeded in developing a model that is able to distinguish between hospitals on the basis of risk structure of their patient populations. We may proceed with analysis.

Figure 46 Differences in "riskiness" of patient populations between hospitals – mean predicted probability of death in a hospital.

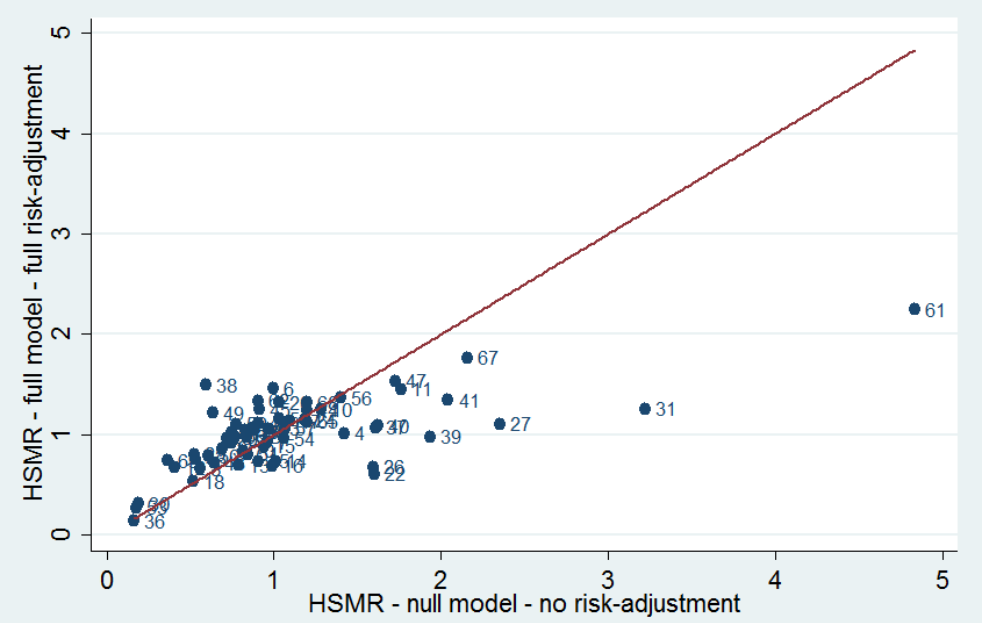
Red line represents mean risk across all hospitals.



Moving on to evaluation of changes in ranking of the hospitals before and after introduction of risk-adjustment, Figure 46 offers a graphical visualization of the changes in calculated hospital mortality ratios after introduction of full risk-adjustment. Red line represents situation when HSMRs from full model are equal to HSMR from null model. At the first glance, hospitals located to the right of the y=x axis exhibit higher deviations than hospitals located to the right. It seems that in the instances when the HSMRs calculated from the null model are underestimated (indicating hospitals that benefit from favorable structure of patients), deviation is higher than in the situation when HSMRs are increased after risk adjustment. However, this plot is purely illustrational as no confidence intervals are considered.

Figure 47. Scatter plot of hospital standardized mortality ratios: full vs null model

Comparison of results from the full model (mortality prediction model) and from the null model (no risk-adjustment, every patient has the same chance of dying)



Moving on to more in-depth analysis, Figure 48 summarizes changes in the classification of hospitals in terms of in-hospital mortality as a quality indicator after introduction of full risk-adjustment (and restricted confidence intervals). Counter-intuitively, only those hospitals with lower CI bound above 1 are considered as below-average and only those hospitals with upper CI bound below 1 are classified as above-standard. When considering only the simplest possible risk-adjustment taking into account only sex and age of the patients, 29.5% of the hospitals are re-classified from one group to another.

If we implement the full model as developed in the previous sections, nearly 50% of all the hospitals are classified in another group – see Figure 49. Admittedly, radical changes from „BAD“ to „GOOD“ group or reversely are rare (only 2 hospitals), still 50% rate of change in classification is substantial. Out of these 19.7% were upgraded (i.e. good=>average; good=>bad; average=>bad) and 26.2% were downgraded. Most misclassification occurred in the identification of above-average hospitals – null model with no risk-adjustment identified 27 hospitals as „GOOD“ and 48% of those were re-classified after introduction of risk-adjustment suggesting that those hospitals undeservedly benefit from favorable structure of patients. Out of those hospitals labelled as sub-standard by the null model, 50% were reclassified as either good or

average after risk-adjustment. Overall, when we compare only number of hospitals in individual categories it becomes obvious that after risk-adjustment model yields more average hospitals: 30 instead of 17 and 12 of those were classified as average by both models. Consequently numbers of „good“ and „bad“ hospitals had to decrease: in the above-average category numbers decreased from 27 to 20 (and 14 were agreed upon by both models) and number of sub-standard hospitals fell down from 14 to 11 (of which 7 were consistently labelled by both models).

Figure 48. Comparison of classification of hospitals between the null model with no risk-adjustment and full model with risk-adjustment.

Classification based on in-hospital mortality

		NULL model - no risk adjustment			
		good	average	bad	total
FULL MODEL	good	14	5	1	20
	average	12	12	6	30
	bad	1	3	7	11
	total	27	20	14	61

** red color - downgrade - 26.2% of all hospitals were downgraded*

blue color - no change - 54.1% of all hospitals recorded no change with risk-adjustment

green color - upgrage - 19.7% of hospitals were upgraded

6.3.2. Comparison of different mortality measures on hospital classification

Finally, we will look at differences in classification of the hospitals when different measures of mortality are used: in-hospital mortality, 30-day mortality and 90-day mortality.

The effect of using different mortality measures is not negligible: 39 hospitals (63%) were classified into the same group by all three mortality indicators, 10 hospitals (16.4%) were classified simultaneously by 30-day and 90-day mortality, 9 hospitals (14.8%) were agreed upon by in-hospital and 30-day mortality and 2 were classified into the same group by in-hospital and 90-day mortality while 30-mortality disagreed. For full classification of hospitals by various mortality indicators see Appendix 10.

Figure 50 presents results for Pearson correlation coefficients between classifications of hospitals with different mortality measures. Correlation between in-hospital

mortality and 30-day mortality and between 30-day and 90-day mortality is relatively high at approximately 0.8, but correlation between in-hospital mortality and 90-day mortality is much weaker – only 0.617. Apparently, choice of mortality indicator does have important consequences for the hospital profiling.

Figure 49. Pearson correlation coefficient between classification of hospitals using measures of in-hospital, 30-day and 90-day mortality

	_INhos~t	_30day~t	_90day~t
_INhospita~t	1.0000		
_30day_mort	0.8048 0.0000	1.0000	
_90day_mort	0.6717 0.0000	0.8097 0.0000	1.0000

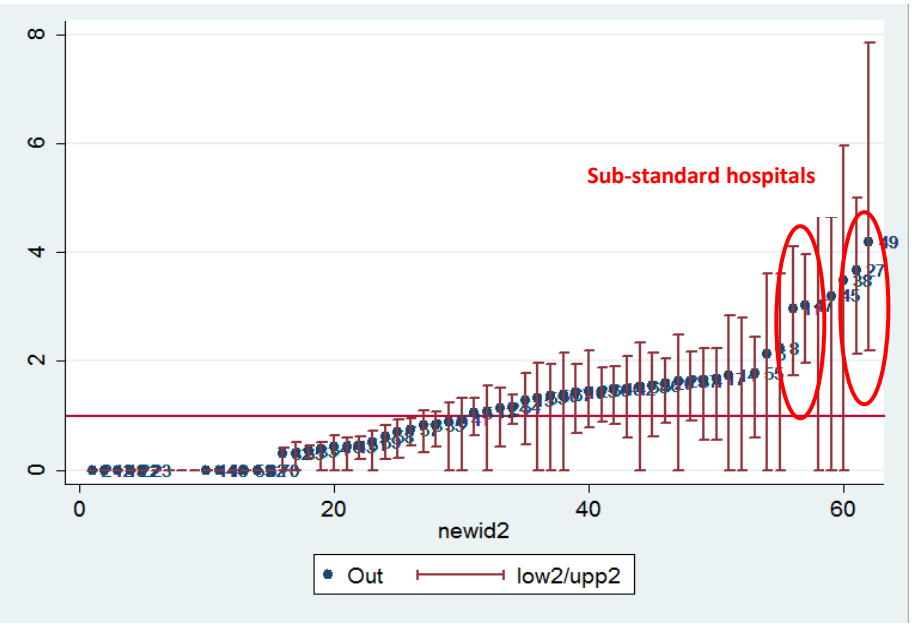
At this stage, we are unable to conclude that one of the mortality measures is any better than another one – we considered statistical properties of the mortality prediction models in the previous section, we looked at the classification outcomes and no measure seems better than another with the exception of 30-day mortality that stands somewhere in between in-hospital mortality and 90-day mortality in terms of classification correlation. Two options appear as a reasonable solution: we could either take into account all three measures - assigning hospital to the category agreed upon by at least two mortality indicators or we could only consider 30-day mortality. Based on the results of correlation analysis, it is not particularly surprising that the results of these two methods are nearly identical: only 2 hospitals are classified differently (change from bad to average and from average to bad). **Simply on the grounds of convenience, we recommend using 30-day mortality for the purposes of hospital profiling** – it is easier to calculate than calculating all three measures. This approach resulted in **12 above-average hospitals (34.4%), 28 average hospitals (45.9%) and 21 below-average hospitals (19.7%).** We verified one more time impact of risk-adjustment on the classification of hospitals, this time 30-day mortality – as the measure selected for the profiling criterion, was compared before and after risk-adjustment. Results are basically identical to the results for in-hospital mortality – 43% of hospitals were reclassified.

6.3.3. Validation of the final model of hospital classification

While my analysis demonstrated that application of risk-adjustment makes a difference – 43% of hospitals are re-classified after we take into account patient-mix treated by the hospitals, we would like some further validation of our results. Since no external validation is available (the only meaningful attempt at hospital profiling by private insurance company is three years old), we considered mortality in low-risk diagnosis categories. Reason why this indicator was not considered along the other three mortality indicators is that it accounts for a very small percentage of overall deaths. While low low-risk mortality (even below-average) is not necessarily an indicator of a good hospital, bad performance is highly probably a sign of low quality of care as it indicates concerns regarding patient safety. Therefore, mortality in high-risk categories is used for validation of the hospital profiling methodology. Should we find out that hospital classified as above-average has significant high-mortality in low-risk diagnosis, we would need to reconsider.

Classification of hospitals was constructed based on HSMRs for low-risk diagnosis, due to low numbers of deaths we risk-adjusted only for age, sex and Charlson comorbidity score and compared results with classification based on 30-day mortality. Only four hospitals are flagged as sub-standard, which is caused by very wide confidence intervals – see Figure 51. This is not surprising given that there are very few deaths in low-risk categories. Out of these four hospitals, our final classification identified three as sub-standard and one as average. But even though our methodology failed to identify this last hospital as sub-standard, it is still an improvement to model with no risk-adjustment that classified this hospital as above-average.

Figure 50. Indicator of low-risk mortality: calculated HSMRs with confidence intervals for hospitals with >100 episodes categorized as low-risk.



7. CONCLUSION

The aim of this thesis was to explore options for hospital profiling in the Slovak republic. Sacrificing breadth of the study in favor of depth, the scope of the analysis was narrowed down to one quality indicator only – mortality. I sought to implement standard techniques used for the purposes of hospital quality assessment in particular conditions of Slovak healthcare systems, necessitating looking for proxy indicators due to local particularities of claims data. Quality profiling is usually a two-step endeavor – in the first step a mortality prediction model is constructed in order to predict expected mortality in hospitals based on a reference sample – usually national average. Predicted risk of mortality is calculated for every single hospital episode on the basis of risk factors that are believed to increase chances of adverse outcome. This process is also called risk-adjustment – taking into account particular case-mix of episodes treated in the hospital. Based on the evidence of best practice, several risk factors were tested in a logistic regression model and what we believe to be a workable mortality prediction model was selected. Risk factors include multiple dimensions that can describe patient populations in individual hospitals including demographic factors, main diagnosis and comorbidities of the patient at the day of admission to the hospital, severity of the episode and past use of healthcare services.

Validation of the model was performed using split-sample approach: model was developed on a 75% derivation sample of the data and then tested on the remaining 25% validation sample. While the calibration of the model was far from perfect – we managed to correctly predict only about 8% of deaths, discriminative ability of the model was exceedingly good – with c-statistics over 0.9 our model is able to distinguish between the episodes with and without adverse outcome exceptionally well. Therefore, while the model is probably not adequate for mortality prediction at individual level, it should perform very well at the aggregated level thanks to its high discriminative ability – such as aggregation at the level of hospitals. Nevertheless, while we consider performance of the model adequate for the purposes of hospital profiling, it might not be a wasted effort to construct mortality prediction coefficients on the bases of repeated random sub-sampling of data – so-called cross-validation.

We also attempted to compare several mortality indicators, including in-hospital mortality, 30-day and 90-day mortality and more specific mortality measures, such as mortality in high-risk conditions or mortality after surgery, but based on the comparison of various statistical properties of models, hospital-wide mortality (regardless of the time frame for observation period) seemed as the best option.

Eventually, 30-day mortality indicator was selected for construction of hospital profiling.

Once the mortality prediction model is developed, we may proceed to second stage – hospital profiling. First of all, we verified our original assumption that motivated this research in the first place – risk-adjustment is crucial for hospital profiling because different hospitals have different risk structure of treated patients. Indeed, mean predicted probability of dying differed significantly between hospitals from 0.2% to 33%. Standard methodological approach commonly used for this type of tasks was then employed: standardized mortality ratios computed as a difference between observed and expected number of deaths, where the latter is calculated from the mortality prediction model. In order to deal with the problem of pure chance and small numbers, we explored options for construction of confidence intervals. In the end, technique of Monte-Carlo simulations was selected – repeating hospital profiling on random sub-samples of data. Single restriction was imposed on the process of generating random sub-samples - mortality in the sub-samples has to be identical to the overall mortality in the data. This approach yielded the widest estimates of confidence intervals, resulting in a relatively strict criterion for identification of outliers – hospitals that can be said that at certain confidence level are different from average. Selection of the computation method of confidence intervals provides a good opportunity for setting the „strictness“ of the profiling. For example, should we calculate confidence intervals using Byar's approximation, we would classify many more hospitals as non-average. Regarding the role of statistical uncertainty, further research should be done, for example calculation of control limits rather than confidence intervals and construction of funnel plots that are appropiate for modelling the effects of randomness. Similarly, techniques of multilevel modelling that separate fixed and random effects should be explored. These techniques are specifically designed to deal with the issue interdependence of within hospitals and are able to estimate betweem-hospital variotion beyond that of a pure chance. Standard models, such as logistic regression used in this analysis assume that individual episodes treated by the hospital are independent, which is highly unlikely.

In the last part of the analysis we looked at the results of profiling to see whether risk-adjustment matters. Comparing classification of hospitals into three groups – above-average, average and substandard we compared results obtained from the model without any risk factors and model including the full risk-adjustment model. 43% of hospitals were re-classified after introduction of risk-adjustment. Acknowledging that different indicators of mortality with full risk-adjustment still yield different results in

terms of hospital categorization with Pearson correlation coefficients between 0.67 and 0.81, we selected 30-day mortality on the criterion of parsimony. Basically identical results can be obtained if we implement a simple algorithm that considers all three mortality measures: hospital is classified into the category agreed upon by at least two of these indicators. Attempt at validation of our results was conducted using the measure of low-risk mortality as a measure of patient safety – negative results are indicative of low quality with a high probability. Our model correctly identified three of those as below average and the remaining hospital was re-classified as average from above-average classification yielded by model without risk-adjustment. However, further validation of the method is needed, for example by conducting randomization tests – comparing actual mortality of the provider against the distribution of mortalities calculated from random samples of episodes with similar patient structure in terms of relevant risk factors to the actual patient mix of the provider.

Therefore, this thesis should be considered as a preliminary analysis of options for hospital profiling in Slovak republic using risk-adjustment to accomodate differences in patient populations treated by individual providers and limited effort was made to eliminate the effects of statistical uncertainty. While the conclusion is straighforward – significant differences in patient populations between hospitals exist and have serious implications for hospital ranking, further research is needed to validate our risk-adjustment model and explore more advanced techniques for dealing with issues of grouped data and randomness.

Bibliography

AHRQ (2008) - Agency for Heathcare Research and Quality: *New AHRQ Study Finds Surgical Errors Cost Nearly $1.5 Billion Annually*. Available online at <http://www.ahrq.gov/news/press/pr2008/surgerrpr.htm>

BAKER, D. – et al. (2002): *Mortality Trends During a Program that Publicly Reported Hospital Performance*, Medical Care, Vol.40, pp. 879-90.

BAKER, D. et al (2003): *The Effect of Publicly Reporting Hospital Performance on Market Share and Risk-Adjusted Mortality at High-Mortality Hospitals*, Medical Care, Vol. 41, Issue 6, pp. 729-740.

BURNS, R.P. – BURNS, R. (2009): *Business Research Methods and Statistics Using SPSS.* Chapter 24: Logistic regression. London: SAGE. Available online at < http://www.uk.sagepub.com/burns/website%20material/Chapter%2024%20-%20Logistic%20regression.pdf>

CHARLSON, M.E. – POMPEI, P. – ALES, K.L. – MACKENZIE, C.R. (1987): *A New Method of Classifying Prognostics Comorbidity In Longitudinal Studies: Development and Validation*, Journal of Chronic Disease, Vol. 40, No.5, pp. 373-383.

CERRITO, P. (2010): *Text Mining Techniques for Healthcare Provider Quality Determination. Methods for Rank Comparison*. New York: Medical Information Science Reference.

CHASSIN, M.R. (2002): *Achieving and Sustaining Improved Quality: Lessons From New York State and Cardiac Surgery.* Health Affairs (Millwood), Vol. 21, No. 4, pp. 40-51.

CIHI (2012) – Canadian Institute for Health Information: *Hospital Standardized Mortality Ratio. Technical notes. Updated April 2012*. Available online at < *Hospital Standardized Mortality Ratio. Technical notes*.>

CUTLER, D.M. – HUCKMAN, R.S. – LANDRUM, M.B. (2004): *The Role of Information in Medical Markets: an Analysis of Publicly Reported Outcomes in Cardiac Surgery*, National Bureau of Economic Research Working Paper 10489. NBER.

D'HOORE, W. – BOUCKAERT, A. – TILQUIN, CH. (1996): *Practical Considerations on the Use of the Charlson Comorbidity Index with Administrative Data Bases*, Journal of Clinical Epidemiology, Vol. 49, No. 12, pp. 1429 – 1433.

Dr. Foster Intelligence (2011): *Hospital Guide 2011*, Available online at < http://drfosterintelligence.co.uk/wp-content/uploads/2011/11/Hospital_Guide_2011.pdf>

DONABEDIAN, A. (1988): *The Quality of Care. How Can it be Assessed*?, The Journal of the American Medical Association, Vol. 260, No. 12, pp. 1743-1748.

EDGMAN-LEVITAN, S. – CLEARY, P. (1996): *What information do consumers want and need?*, Health Affairs, Vol. 15, pp. 42-56.

GRUMBACH, K. et al. (1998): *Primary Care Physicians' Experience of Financial Incentives in Managed-Care Systems*, The New England Journal of Medicine, Vol. 339, pp. 1516-1521.

HAFNER, J.M. – WILLIAMS, S.C. – KOSS, R.G. – TSCHURTZ, B.A. – SCHMALTZ, S.P. – LOEB, J.M. (2011): *The Perceived Impact of Public Reporting Hospital Performance Data: Interviews With Hospital Staff*, International Journal for Quality In Healtcare, Vol.23, No.6, pp.697-704.

HANNAN, E.L. – KILBURN, H. – RACZ, M. – SHIELDS, E. – CHASSIN, M.R. (1994): *Improving the Outcomes of Coronary Artery Bypass Surgery in New York State*, Journal of the American Medical Association, Vol. 271, pp. 761-66.

HANNAN, E. L. et al (2003): *Do Hospitals and Surgeons with Higher Coronary Artery Bypass Graft Surgery Volumes Still Have Lower Risk-Adjusted Mortality Rates?* Circulation, Vol. 108, No. 7, pp. 795-801.

HIBBARD, J. – STOCKARD, J. – TUSLER, M. (2005): *Hospital Performance Reports: Impact on Quality, Market Share, and Reputation*, Health Affairs (Millwood), Vol. 24, No. 4, pp. 1150-1560.

HIBBARD, J. – STOCKARD, J. – TUSLER, M. (2003): *Does Publicizing Hospital Performance Stimulate Quality Improvement Efforts?*, Health Affairs (Millwood), Vol. 22, No. 2, pp. 84-94.

HIBBARD, J. – SOFAER, S. – JEWETT, J. (1996): *Condition-specific Performance Information: Assessing Salience, Comprehension and Approaches for Communicating Quality*, Health Care Financing Review, Vol. 18, No. 1, pp. 95-109.

HIBBARD, J. – JEWETT, J. (1997): *Will Quality Report Cards Help Consumers?*, Health Affairs, Vol. 16, pp. 218-228.

HIBBARD, J. – SOFAER, S (2010): *Best Practices in Public Reporting No. 1. - How To Effectively Present Health Care Performance Data To Consumers*, AHRQ Publication No. 10-0082-EF. Available online at < http://www.ahrq.gov/qual/pubrptguide1.htm >

IDRE (2012): Online lecture notes. Institute For Digital Research and Education.. Available at <http://www.ats.ucla.edu/stat/stata/webbooks/logistic/chapter3/statalog3.htm>

IOM (2000) – Institute of Medicine: *To Err is Human*. Available online at <http://iom.edu/~/media/Files/Report%20Files/1999/To-Err-is-Human/To%20Err%20is%20Human%201999%20%20report%20brief.pdf>

ITSKOVITSCH, I. – ROUDEBUSH, B. (2010): *Using Re-Sampling Methods in Mortality Studies*. PLoS ONE, Vol. 5, No. 8. Available online at < http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0012340>

LINGSMA, H.F. – STEYERBERG, E.W. – EIJKEMANS, M.C.J. - D.W.DIPPEL, REIMER, W.J.M. – HOUWELINGEN, H.C.Van (2009): *Comparing and Ranking Hospitals Based on Outcome: Results from the Netherlands Stroke Survey*, QJM – International Journal of Quality, Vol. 103, No.2, pp. 99-108.

LONGO, D.R. – LAND, G. – SCHRAMM, W. – FRAAS, J. – HOSKINS, B. – HOWELL, V. (1997): *Consumer Reports in Health Care. Do They Make a Difference in Patient Care?*, Journal of the American Medical Association, Vol. 278, pp. 1579-1584.

LUCE, J.M. – THIEL, G.D. - HOLLAND, M.R. - SWIG, L. - CURRIN, S.S. – LUFT, H.S. (1996): *Use of Risk-adjusted Outcome Data for Quality Improvement by Public Hospitals*, Western Journal of Medicine, Vol. 164, pp. 410-414.

LUNT, M. (2012): *Modelling Binary Outcomes.* Available at <http://personalpages.manchester.ac.uk/staff/mark.lunt/stats/7_Binary/text.pdf>

KANSAGARA, D. – ENGLANDER, H. – SALANITRO, A. – KAGEN, D. – THEOBALD, C. – FREEMAN, M. – KRIPALANI, S. (2011): Risk Prediction Models for Hospital Readmission. A Systematic Review, Journal of the American Medical Association, Vol. 306, No.15, pp. 1688-1689.

LANSKY, D. (2002): *Improving quality through public disclosure of performance information*, Health Affairs, Vol. 21, Issue 4, pp. 52-62.

MANNION, R. – GODDARD, M. (2003): *Public disclosure of comparative clinical performance data: lessons from the Scottish experience*, Journal of Evaluation in Clinical Practice, Vol. 9, No. 2, pp. 277-286.

MARSHALL, M. N. et al. (2000): *The Public Release of Performance Data. What Do We Expect to Gain? A Review of the Evidence*, The Journal of the American Medical Association, Vol. 283, No. 14, pp. 1867-1874.

MENNEMEYER, S. T. et al (1997): *Death and reputation: how consumers acted upon HCFA mortality information*, Inquiry, Vol. 34, No. 2, pp. 117-128.

NARRINS, C.R. – DOZIER, A.M. – LING, F.S., ZAREBA, W. (2005): *The influence of public reporting of outcome data on medical decision making by physicians,* Archives of Internal Medicine, Vol. 165, pp. 83-87.

New York Times (2007): *N.Y. Attorney General Objects to Insurer's Ranking of Doctors by Cost and Quality*. Newspaper article, published July, 14th 2007. Available online at <http://www.nytimes.com/2007/07/14/nyregion/14healthcare.html?_r=0>

NHS (2012): *Indicator Specification: Summary Hospital-level Mortality Indicator*, Available online at < http://www.ic.nhs.uk/CHttpHandler.ashx?id=10628&p=0>

NORMAND, S-L.T – SHAHIAN, D.M. (2007): *Statistical and Hospital Aspects of Hospital Outcomes Profiling, Statistical Science*, Vol.22, No.2, pp. 206-226.

PEDUZZI, P. – CONCATO, J. – KEMPER, E. – HOLFORD, T.R. – FEINSTEIN, A.R. (1996): *A simulation study of the number of events per variable in logistic regression analysis*, Journal of Clinical Epidemiology, Vol. 49, No. 12, pp. 1373 - 1379.

PENG, Ch.-Y. J. – SO, T.-S. H. (2002): *Logistic Regression Analysis and Reporting: A primer*, Teaching article. Available online at < http://www.indiana.edu/~jopeng51/teaching-logistic.pdf>

ROBINSON, S. – BRODIE, M. (1997): *Understanding the quality challenge for health consumers: the Kaiser/AHCPR survey*, Joint Commission Journal on Quality Improvement, Vol. 23, pp. 239-244.

ROBINSON, S. – MOLLYANN, B. (1997): *Understanding the quality challenge for health consumers: the Kaiser/ AHCPR survey*, Joint Commission Journal for Quality Improvement, Vol. 23, pp. 239-244.

ROMANO, P. S. – ZHOU, H. (2004): *Do Well-Publicized Risk-Adjusted Outcomes Reports Affect Hospital Volume?* Medical Care, Vol. 42, No. 4, pp. 367–77.

ROSENTHAL, M. B. et al. (2005): *Early Experience With Pay-for-Performance From Concept to Practice*, The Journal of the American Medical Association, Vol. 294, No. 14, pp. 1788-1793.

SCHNEIDER, E. C. – EPSTEIN, A. M. (1998): *Use of Public Performance Reports. A Survey of Patients Undergoing Cardiac Surgery*, The Journal of the American Medical Association, Vol. 279, No. 20.

SCHWARTZ, M. – ASH, A. – PEKÖZ, E. (2006): *Risk adjustment and risk-adjusted provider profiles*, International Journal of Heathcare Technology and Management, Vol. 7, Nos. 1/2.

SHAHIAN, D.M. – IEZZONI, L.I. – MEYER, G.S. – KIRLE, L. – NORMAND, S.L. (2012) *Hospital-wide Mortality as a Quality Metric: Conceptual and Methodological Challenges*, American Journal of Medical Quality, Vol. 27, No. 2, pp. 112-123.

SHAHIAN, D.M. – WOLF, R.E. – IEZZONI, L.I. – LIRLE, L. – NORMAND, S-L.T. (2010): *Variability in the Measurement of Hospital-wide Mortality Rates*, The New England Journal of Medicine, Vol. #29, No. 26, pp. 2530 -2539.

SHEKELLE, P. G. et all (2008): *Does public release of performance results improve quality of care?* The Health Foundation, Available online at < http://www.health.org.uk/public/cms/75/76/313/554/Public%20release%20of%20performance%20result.pdf?realName=UWXIXp.pdf>

SOE, M.M. – SULLIVAN, K.M. (2006): *Standardized Mortality Ratio and Confidence Interval.* Available online at < http://www.openepi.com/documentation/smrdoc.htm>

STAUSBERG, J. – HALIM, A. – FÄRBER, R. (2011): *Concordance and Robustness of Quality Indicator Sets for Hospitals: an Analysis of Routine Data*, BMC Health Services Research, Vol. 11. Available online at < http://www.biomedcentral.com/1472-6963/11/106>

STEELE, F. (2010): Online lecture notes: *Introduction to Multilevel Modelling Concepts*. Centre for Multilevel Modelling. University of Bristol. Available online (after registration): < http://www.cmm.bris.ac.uk/lemma/course/view.php?id=13>

TU, J.V – DONOVAN, L.R – LEE, D.S. (2009): *Effectiveness of Public Report Cards for Improving the Quality of Cardiac Care.* The EFFECT Study: A randomized trial. *JAMA.* Available online at < http://www.theheart.org/article/1026701.do>

TU, J.V. – SYKORA, K. – NAYLOR. C.D. (1997): *Assessing the Outcomes of Coronary Artery Bypass Graft Surgery: How Many Risk Factors are Enought?*, Journal of the America College of Cardiologists, Vol. 30, No.5, pp. 1317-1323.

UDZS (2011) – Vestník č. 7/2011. Správa o činnosti Úradu za rok 2010. Available online http://www.udzs-sk.sk/buxus/docs//vestniky/rocnik_2011/VE_7_Sprava_cinn_tlac.pdf

UNIVERSITY of STRATHCLYDE (2012): online lecture notes on logistic regression. Available online at <http://www.strath.ac.uk/aer/materials/5furtherquantitativeresearchdesignandanalysis/unit6/goodnessoffitmeasures/>

VAINA, M. – MCGLYNN, E. (2002): *What Cognitive Science Tells us About the Design of Reports for Consumers*, Medical Care Research and Review, Vol. 59, No. 1, pp. 3-35.

WERNER, R.M. – ASCH, D.A. (2005): *The Unintended Consequences of Publicly Reporting Quality Information.* Journal of the American Medical Association, Vol. 239, pp. 1239-1244.

APPENDIX 1: List of information in claims data

Compulsory data for inpatient care insurance claims include following information (but of course, few of these are reported consistently):

- Identification number of the patient, of the provider organization and doctor
- Name of the patient
- Diagnosis code at the admission (ICD-10 classification)
- Admission date
- Discharge date
- Number od days (length of stay)
- Movement of the patient (transfer to other hospital, transfer to other department in the same hospital, transfer to specialized medical institution, discharge home, transfer to social care institution, death, discharge at patient's request)
- Newborn
- Additional items (not covered by the flat rate case payment, are reimbursed in addition to the case rate)
- Amount (of additional items)
- Price (of additional items)
- Compensations (to be claimed by the insurance company: occupation accident, ...)
- State of the insuree (N – normal, I – complicated, 3 – extremely complicated)
- Type of procedure (O – operation, S – intervention radiology, N – none)
- Code of operation
- Code of acquired complication (O-without complication, 1 – nosocomial infection, 2 – decubity, 3 – digestion problems, 4 – pneumony, 5 – other, 6 – 2 and more complications)
- Type of hospitalization (A – acute, C – centrally purchased drugs, E – planned operation, P – planned operations included on the official waiting lists, I – other, K – spa treatment, S – day care centre, Z – additional item)
- Comorbidity
- Admission with complication
- Date of admission
- Date od discharge

APPENDIX 2: High-risk conditions

Diagnosis category	Mortality	Diagnosis name	Diagnosis codes
DGN001	14,1%	Malignant neoplasm of oesophagus, stomach and pancreas	C15, C16, C25
DGN002	10,6%	Malignant neoplasms of respiratory and intrathoracic organs	C32, C34
DGN003	76,7%	cardiac arrest	I46
DNG004	13,5%	heart failure	I50
DGN005	35,9%	acute respiratory failure	J96, J95
DGN006	12,2%	Pulmonary oedema	J81, J90, J91
DGN007	12,9%	diseases of liver	K70, K71, K72, K74, K76
DGN008	16,0%	disorders of fluid, electrolyte and acid-base balance	E86, E87
DGN009	10,1%	Injuries involving multiple body sites (polytrauma)	T06, T07
DGN010	12,5%	Other malignant neoplasms	C04, C09, C10, C13, C22, C23, C24, C26, C32, C49, C71, C79, C80
DGN011	11,6%	Other cardiovascular disease	I42, I95, I97
DGN012	10,2%	Other cerebrovascular disease	G45, I60, I61, I63, I64, I67, I69

APPENDIX 3: Low-risk diagnosis

Source: Dr Foster (2012a)

Viral infection	A70,A82,A881,A90,A91,A920,A921,A923,A924,A928,A929,A93,A94,A950,A951,A959,A96,A98,A99,B000,B001,B002,B008,B009,B018,B019,B022,B027,B028,B029,B03,B04,B059,B068,B069,B07,B08,B09,B258,B259,B260,B268,B269,B27,B330,B331,B333,B338,B34,B97,P352,U04
Sexually transmitted infections (not HIV or hepatitis)	A50-A60,A63,A64,I980,N742-N744
Immunizations and screening for infectious disease	R761,R762,Z030,Z11,Z20,Z22-Z27
Benign neoplasm of uterus	D25,D26
Other and unspecified benign neoplasm	D10-D24,D27-D36
Immunity disorders	D80-D84,D89
Sickle cell anaemia	D57
Other haematologic conditions	D730-D732, D734,D735,D738,D739,D74,D75,D77,R71
Mental retardation	F70-F73,F78,F79,F844
Alcohol-related mental disorders	F10,G312,R780
Substance-related mental disorders	F11-F19,F55,R781-R784
Anxiety, somatoform, dissociative, and personality disorders	F40,F410,F411,F413,F418,F419,F42-F45,F48,F51,F60-F63,F68,F69,R451
Preadult disorders	F840-F843,F845,F848,F849,F90-F94,F98
Headache, including migraine	G43,G44,R51
Cataract	H25,H26,H280-H282
Retinal detachments, defects, vascular occlusion, and retinopathy	G453,H33-H36
Glaucoma	H40,H420,H428
Blindness and vision defects	H52-H54,H581

Inflammation, infection of eye	A211,A71,A74,B005,B023,B058,B30,B580,B601,B691,B872,B940,H000,H01,H03,H040,H043,H044,H050,H051,H061,H10,H13,H150,H151,H161-H163,H168,H169,H190-H193,H20,H220,H221,H30,H320,H440,H441,H451,H46,H481
Other eye disorders	H001,H02,H041,H042,H045-H049,H052-H060,H062,H063,H11,H158-H160,H164,H17,H18,H198,H21,H228,H27,H288,H31,H328,H43,H442-H450,H458,H47,H480,H488,H49-H51,H55,H57,H580,H588
Otitis media and related conditions	B053,H65-H70,H72,H73,H740-H743,H75,H80
Conditions associated with dizziness or vertigo	H81-H83,R42
Other ear and sense organ disorders	B874,H60-H62,H71,H744,H748,H749,H90-H93,H940,H948
Nonspecific chest pain	R071-R074
Varicose veins of lower extremity	I83
Haemorrhoids	I84
Acute and chronic tonsillitis	J03,J35,J36
Other upper respiratory infections	A360-A362,A37,B873,J00-J02,J04-J06,J32
Asthma	J45,J46
Disorders of teeth and jaw	K00-K08,K090-K092,K10
Diseases of mouth, excluding dental	A690,A691,K098,K099,K11-K14,R682
Oesophageal disorders	I859,I982,K20-K23
Gastritis and duodenitis	K29
Other disorders of stomach and duodenum	K30,K31
Appendicitis and other appendiceal conditions	K35-K38
Abdominal hernia	K40-K46
Anal and rectal conditions	K594,K60,K61,K620-K624,K626-K629
Calculus of urinary tract	N20,N21,N220,N228,N23

Other diseases of bladder and urethra	N31,N32,N338,N350,N358,N359,N36
Genitourinary symptoms and ill-defined conditions	N02,N391,N393,N394,N398,N399,R30-R36,R39,R80,R820,R821-R823,R825-R829,R934,R944
Hyperplasia of prostate	N40
Inflammatory conditions of male genital organs	N41,N431,N45,N482,N486,N49,N51
Other male genital disorders	N42,N430,N432-N434,N44,N46,N47,N480,N481,N483-N485,N488,N489,N50,R86
Nonmalignant breast conditions	N60-N64
Inflammatory diseases of female pelvic organs	A483,N70-N73,N748,N750,N751,N76,N77
Endometriosis	N80
Prolapse of female genital organs	N81
Menstrual disorders	N91,N92,N938,N939,N944-N946
Ovarian cyst	N830-N832
Menopausal disorders	N95
Female infertility	N97
Other female genital disorders	N758,N759,N82,N833-N839,N84-N90,N930,N940-N943,N948,N949,N96,R87
Contraceptive and procreative management	Z30,Z31,Z320,Z35
Spontaneous abortion	O03
Induced abortion	O04-O07
Postabortion complications	O08
Ectopic pregnancy	O00
Other complications of pregnancy	O01,O02,O12,O21,O23,O25,O260-O264,O266-O269,O28,O31,O860-O863,O98,O99
Haemorrhage during pregnancy, abruptio placenta, placenta previa	O20,O44-O67

Hypertension complicating pregnancy, childbirth and the puerperium	O10,O11,O13-O16
Early or threatened labour	O47
Prolonged pregnancy	O48
Diabetes or abnormal glucose tolerance complicating pregnancy, childbirth, or the puerperium	O24
Malposition, malpresentation	O32,O64,O801,O830,O831
Fetopelvic disproportion, obstruction	O33,O65,O66
Previous C-section	O757
Foetal distress and abnormal forces of labour	O363,O62,O63,O68
Polyhydramnios and other problems of amniotic cavity	O40-O42,O755,O756
Umbilical cord complication	O69
Trauma to perineum and vulva	O70
Forceps delivery	O81,O841
Other complications of birth, puerperium affecting management of mother	A34,O22,O265,O34,O35,O360-O362,O364-O369,O43,O60,O61,O71-O74,O750-O754,O758,O759,O82,O832-O834,O838,O839,O842,O848,O85,O864,O868,O87,O88,O90-O92,O95-O97
Normal pregnancy and/or delivery	O30,O800,O808,O809,O840,O849,Z321,Z33,Z34,Z37,Z39
Other inflammatory condition of skin	L10,L12-L14,L21,L26,L28,L29,L304,L305,L308,L309,L40-L42,L430,L431,L433,L438-L443,L448,L449,L45,L510,L511,L518,L519,L52,L531-L533,L538-L540,L548,L661,L71,L920,L93,L945,L951,L981,L982
Other skin disorders	L11,L301,L57,L60,L62,L63,L648,L649,L65,L660,L662,L663,L664,L668,L669,L67,L68,L70,L72-L75,L80-L87,L90,L91,L921-L923,L928,L929,L940-L944,L948-L950,L958,L959,L985-L989,L99,R21,R22,R234,R238,R61
Rheumatoid arthritis and related disease	M05,M06,M08,M09,M120

Osteoarthritis	M15-M19
Acquired foot deformities	M201-M206,M214,M216
Other acquired deformities	M200,M210-M213,M215,M217-M219,M245,M40,M430,M431,M438,M439,M95
Other connective tissue disease	M242,M257,M353-M357,M60-M62,M630-M633,M638,M65-M79,R293,R298,R936,R937
Other bone disease and musculoskeletal deformities	M41,M42,M840-M842,M848,M849,M85,M870,M88,M89,M906,M908,M91-M94,M99
Digestive congenital anomalies	Q38-Q45
Genitourinary congenital anomalies	Q50-Q64
Other congenital anomalies	Q10-Q18,Q30-Q37,Q65-Q99,R294
Liveborn	Z38
Intrauterine hypoxia and birth asphyxia	P20,P21
Haemolytic jaundice and perinatal jaundice	P546-P549,P55-P59
Birth trauma	P10-P15
Joint disorders and dislocations, trauma-related	M125,M22,M23,M241,M244,S030-S033,S130-S133,S230-S232,S330-S334,S430-S433,S530-S533,S630-S634,S730,S830-S833,S837,S930-S933,T03,T092,T112,T132,T143,T923,T933
Sprains and strains	S034,S035,S134-S136,S233-S235,S335-S337,S434-S437,S534,S635-S637,S731,S834-S836,S934-S936
Open wounds of extremities	S411,S51,S58,S61,S68,S711,S781,S789,S81,S88,S91,S98,T012,T013,T016,T050-T056,T111,T116,T131,T136,T920,T930
Poisoning by psychotropic agents	T40,T420-T427,T43
Poisoning by other medications and drugs	L640,N14,T36-T39,T41,T428,T44-T50,T96
Lymphadenitis	I88,L04,R59
Abdominal pain	R10

Allergic reactions	L20,L22-L25,L27,L300,L302,L432,L50,L512,L530,L55,L56,L58,L59,T780-T784
Medical examination/evaluation	Z00,Z01,Z04,Z10
Other aftercare	Z08,Z09,Z42,Z47,Z48,Z54
Other screening for suspected conditions	Z031,Z033-Z039,Z12,Z130-Z132,Z134-Z139,Z36

APPENDIX 4: List of codes for identification of implantation of artificial joints

product code	product name
130001	Cementovaná TEP bedrového kĺbu
130002	Hybridná TEP bedrového kĺbu
130003	Necementovaná TEP bedrového kĺbu
130009	Cementovaná TEP kolenného kĺbu
130011	Necementovaná unikondylárna TEP kolena
130101	Cementovaná TEP bedrového kĺbu
130201	Hybridná TEP bedrového kĺbu
130301	Necementovaná TEP bedrového kĺbu
130302	Necementovaná TEP s keramickými artikulačnými povrchmi
130303	Necementovaná TEP s kovovými artikulačnými povrchmi
130501	Individuálna necementovaná TEP bedrového kĺbu
130705	Použitie revízneho implantátu pri primárnej indikácii pre TEP kolena
130801	Individuálna TEP kolenného kĺbu
130901	Cementovaná unikondylárna TEP kolenného kĺbu
130902	Cementovaná all-polly TEP kolenného kĺbu
130903	Cementovaná fixná TEP kolenného kĺbu
130904	Cementovaná rotačná TEP kolenného kĺbu
131002	Hybridná fixná TEP kolenného kĺbu
131003	Hybridná rotačná TEP kolenného kĺbu
131101	Necementovaná fixná TEP kolena
131102	Necementovaná rotačná TEP kolena
131201	TEP ramena

APPENDIX 5: Diagnosis categories

Diagnosis categories	Diagnosis name	Diagnosis codes	Number of cases	Number of deaths	Mortality in %
DGN01	Malignant neoplasms of digestive organs	C15, C16, C18, C19, C20, C22, C23, C25	3465	358	10,3%
DGN02	Malignant neoplasms of respiratory and intrathoracic organs	C32, C34	2097	222	10,6%
DGN03	Other neoplasms	C43, C50, C53, C56, C61, C64, C67, C71, C78, C92, D38	5371	339	6,3%
DGN04	Endocrine, nutritional and metabolic diseases	E10, E11, E86, E87	487	453	93,0%
DGN05	Ischemic heart disease	I20, I21, I25	8352	354	4,2%
DGN06	Other heart diseases	I26, I46, I48, I49	5686	303	5,3%
DGN07	Heart failure	I50	3602	494	13,7%
DGN08	Cerebral infarction	I63	3139	316	10,1%
DGN09	Other cerebrovascular disease	I60, I61, I64, I67, I69	428	521	121,7%
DGN10	Influenza and pneumonia	J15, J18	7084	312	4,4%
DGN11	Other diseases of respiratory system	J20, J44, J81, J90, J95, J96	10372	490	4,7%
DGN12	Diseases of liver	K70, K74, K76	2119	256	12,1%
DGN13	Other diseases of digestive system	K30, K56, K80, K83, K85, K	15233	560	3,7%
DGN14	General symptoms and signs	R50, R55, R57, R63	3112	69	2,2%
DGN15	Other symptoms involving specific system	R06, R10, R26, R40	3277	96	2,9%
DGN16	Injuries	S06, S72, T07	7537	311	4,1%
DGN17	Other	A41, I10, G40, I95, D50, D64, N17, N18, I70, I71	12114	457	3,8%

APPENDIX 6: Charlson Comorbidity index conditions

Charlson Comorbidity index conditions (based on ICD-10 codes). New and old weights. Source: NHS (2012)

Condition	Condition Name	ICD-10 codes	New weight	Old weight
1	Acute myocardial infarction	I21, I22, I23, I252, I258	5	1
2	Cerebral vascular accident	G450, G451, G452, G454, G458, G459, G46, I60-I69	11	1
3	Congestive heart failure	I50	13	1
4	Connective tissue disorder	M05, M060, M063, M069, M32, M332, M34, M353	4	1
5	Dementia	F00, F01, F02, F03, F051	14	1
6	Diabetes	E101, E105, E106, E108, E109, E111, E115, E116, E118, E119, E131, E131, E136, E138, E139, E141, E145, E146, E148, E149	3	1
7	Liver disease	K702, K703, K717, K73, K74	8	1
8	Peptic ulcer	K25, K26, K27, K28	9	1
9	Peripheral vascular disease	I71, I739, I790, R02, Z958, Z959	6	1
10	Pulmonary disease	J40-J47, J60-J67	4	1
11	Cancer	C00-C76, C80-C97	8	2
12	Diabetes complications	E102, E103, E104, E107, E112, E113, E114, E117, E132, E133, E134, E137, E142, E143, E144, E147	-1	2
13	Paraplegia	G041, G81, G820, G821, G822	1	2
14	Renal disease	I12, I13, N01, N03, N052-N056, N072-N074, N18, N19, N25	10	2
15	Metastatic cancer	C77, C78, C79	14	3
16	Severe liver disease	K721, K729, K766, K767	18	3
17	HIV	B20, B21, B22, B23, B24	2	6

APPENDIX 7: Adding interaction terms – comparison before and after

Comparison of performance of model after addition of two interaction terms – between age and Charlson score nad between age and high-risk diagnosis

	MODEL3a Age, Sex, Charlson comorbidity Index, TransferIN, emergency transport, emergency ARO, No of previous hospitalizations, **3 diagnosis groups - high risk, medium risk and low risk**	**MODEL4** Age, Sex, Charlson comorbidity Index, TransferIN, emergency transport, emergency ARO, No of previous hospitalizations, **3 diagnosis groups - high risk, medium risk and low risk + 2x interaction terms: age*Charlson, age*DGN_HR**
pseudo R-squared (McFadden's)	0.2917	0.2956
LR chi2(0)	19799.23	20065.28
Prob > chi2	0.000	0.000
C-statistics	0.9027	0.9058
Predictive accuracy	97.57%	97.58%
Sensitivity	8.29%	8.75%
Specificity	99.85%	99.85%
Positive predictive value	59.33%	59.57%
Negative predictive value	97.71%	97.72%
Hosmer-Lemeshow test - **chi-square**	177.64	125.13
Hosmer-Lemeshow test - **p-value**	0.000	0.000

APPENDIX 8: Stability of estimated coefficients across derivation and validation samples

Comparison of odds calculated separately for derivation and validation sample.

LEFT – in-hospital mortality. RIGHT – 30-day mortality

	DERIVATION sample	VALIDATION sample
LC_DeathIN	Odds Ratio	Odds Ratio
LP_Sex	.7553201	.7120394
age1	.4383333	.1033483
age2	1.17824	.7674589
age3	2.835385	2.540376
age4	5.709615	4.894021
age5	9.247014	6.503084
age6	13.93196	10.63033
age7	21.83413	17.46151
age8	48.8141	39.20113
age9	94.02708	72.72165
ch1	1.872452	1.727299
ch2	2.490874	2.113608
ch3	3.879958	3.32713
ch4	6.443838	4.855428
ch5	9.600571	8.151796
LC_TranferIN	2.455285	2.079074
CHAcp2	.59553	.6807404
CHA2c	.8787979	.8175773
LC_EMERG2_~t	2.261421	2.285711
LP_NoHosp2	1.267008	1.248011
DGN_MR	1.937485	2.006729
DGN_HR	4.176285	5.172274
DGN_LR	.4434858	.4701924
LC_EMERG1_~O	13.20509	12.15047

	DERIVATION sample	VALIDATION sample
LC_Death1M	Odds Ratio	Odds Ratio
LP_Sex	.7202157	.6856271
age1	.4296003	.2229755
age2	1.237342	.7113377
age3	2.751971	2.387518
age4	5.800557	5.381395
age5	10.203	7.832974
age6	14.88945	13.34661
age7	23.45212	21.56097
age8	55.62122	51.79928
age9	125.6252	114.0797
ch1	1.900223	1.886287
ch2	2.505174	2.336772
ch3	3.972116	3.818215
ch4	6.21413	5.533616
ch5	8.909203	8.193824
LC_TranferIN	2.234768	1.833514
CHAcp2	.6342632	.6345609
CHA2c	.8535517	.8164347
LC_EMERG2_~t	1.990645	2.172815
LP_NoHosp2	1.349053	1.336587
DGN_MR	1.998834	2.061034
DGN_HR	4.425923	5.007567
DGN_LR	.5082308	.5619194
LC_EMERG1_~O	9.811766	9.82444

APPENDIX 9: Results for mortality prediction model used for hospital profiling

Model used for hospital profiling (period 2011.07-2012.06)

Logistic regression — Number of obs = 147494; LR chi2(24) = 10099.92; Prob > chi2 = 0.0000; Log likelihood = -12731.032; Pseudo R2 = 0.2840

LC_DeathIN	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
LP_Sex	.7303553	.0266651	-8.61	0.000	.6799188	.7845331
age1	.2644847	.1436749	-2.45	0.014	.0912023	.7669999
age2	.9126401	.2682509	-0.31	0.756	.512988	1.623648
age3	2.655422	.6270974	4.14	0.000	1.671537	4.218432
age4	4.530629	1.008071	6.79	0.000	2.929309	7.00732
age5	7.331326	1.565876	9.33	0.000	4.823663	11.14264
age6	11.64118	2.47799	11.53	0.000	7.670204	17.668
age7	18.4424	3.969167	13.54	0.000	12.09545	28.11983
age8	39.56421	8.659984	16.80	0.000	25.76249	60.75991
age9	68.34108	16.14727	17.88	0.000	43.0095	108.5924
ch1	1.705498	.1103999	8.25	0.000	1.502282	1.936203
ch2	2.180057	.1913695	8.88	0.000	1.835472	2.589333
ch3	3.233152	.3989109	9.51	0.000	2.538655	4.117642
ch4	4.793783	.7893929	9.52	0.000	3.47145	6.619815
ch5	7.473495	1.789034	8.40	0.000	4.674766	11.94779
LC_TranferIN	2.497139	.1730739	13.20	0.000	2.179952	2.860478
CHAcp2	.6149689	.0604811	-4.94	0.000	.5071531	.7457054
CHA2c	.860101	.0234319	-5.53	0.000	.81538	.9072749
LC_EMERG2_~t	2.164458	.079708	20.97	0.000	2.013738	2.326458
LP_NoHosp2	1.244021	.032914	8.25	0.000	1.181155	1.310233
DGN_LR	.4768366	.0398904	-8.85	0.000	.404726	.5617952
DGN_MR	2.046155	.104561	14.01	0.000	1.851148	2.261705
DGN_HR	4.774677	.3980693	18.75	0.000	4.054885	5.622241
LC_EMERG1_~O	13.10473	.9237651	36.50	0.000	11.41369	15.04631

Logistic model for LC_DeathIN

Classified	True D	True ~D	Total
+	323	227	550
-	3512	143432	146944
Total	3835	143659	147494

Classified + if predicted Pr(D) >= .5
True D defined as LC_DeathIN != 0

Sensitivity	Pr(+| D)	8.42%
Specificity	Pr(-|~D)	99.84%
Positive predictive value	Pr(D| +)	58.73%
Negative predictive value	Pr(~D| -)	97.61%
False + rate for true ~D	Pr(+|~D)	0.16%
False - rate for true D	Pr(-| D)	91.58%
False + rate for classified +	Pr(~D| +)	41.27%
False - rate for classified -	Pr(D| -)	2.39%
Correctly classified		97.46%

APPENDIX 10: Comparison of classification of hospitals using different measures of mortality.

Green – above average, blue – average, red- below-average. Numbers identify individual hospitals.

In-hospital M NO risk-adjust.	In-hospital M FULL risk-adjust.	30-day M FULL risk-adjust.	90-day M FULL risk-adjust.	MODEL combining all 3 Ms
3	3	3	3	3
5	5	5	5	5
7	7	7	7	7
8	8	8	8	8
12	12	12	12	12
13	13	13	13	13
14	14	14	14	14
15	15	15	15	15
16	16	16	16	16
18	18	18	18	18
22	22	22	22	22
26	26	26	26	26
30	30	30	30	30
33	33	33	33	33
35	35	35	35	35
36	36	36	36	36
43	43	43	43	43
46	46	46	46	46
50	50	50	50	50
53	53	53	53	53
69	69	69	69	69
2	2	2	2	2
4	4	4	4	4
9	9	9	9	9
10	10	10	10	10
23	23	23	23	23
24	24	24	24	24
25	25	25	25	25
29	29	29	29	29
32	32	32	32	32
34	34	34	34	34
37	37	37	37	37
39	39	39	39	39
40	40	40	40	40
42	42	42	42	42
44	44	44	44	44
51	51	51	51	51
54	54	54	54	54
55	55	55	55	55
56	56	56	56	56
57	57	57	57	57
58	58	58	58	58
59	59	59	59	59
62	62	62	62	62
65	65	65	65	65
66	66	66	66	66
68	68	68	68	68
70	70	70	70	70
49	49	49*LR	49	49
6	6	6	6	6
17	17	17	17	17
20	20	20	20	20
31	31	31	31	31
38	38	38	38	38
41	41	41	41	41
45	45	45	45	45
61	61	61	61	61
67	67	67	67	67
11	11	11*LR	11	11
27	27	27*LR	27	27
47	47	47*LR	47	47

**LR - flagged as sub-standard by low-risk mortality indicator*