A quality indicator can be biased by intra-hospital heterogeneity: the case for quality of patient record keeping in France

Audrey Lelong1,2, Stéphanie Léger1,3,4, Françoise Vendittelli1,2,5, Marie Blanquet1, Cong-Tri Thuong1, Bénédicte Belgacem1,2, Laurent Gerbaud1,2

1 Department of Public Health, Clermont-Ferrand University Hospital Center, Clermont-Ferrand, France
2 EA 4681, PEPRADE, University of Auvergne, Clermont-Ferrand, France
3 Laboratoire de Mathématiques, Université Clermont Auvergne, Université Blaise Pascal, BP 10448 F-63000 Clermont-Ferrand, France
4 CNRS, UMR 6620, Laboratoire de Mathématiques, F-63171 Aubière, France
5 Department of Obstetrics & Gynecology, Clermont-Ferrand University Hospital Center, Clermont-Ferrand, France

Correspondence: Audrey Lelong, Department of Public Health, Clermont-Ferrand University Hospital Center, 7 Place Henri Dunant, 63000 Clermont-Ferrand, France, Tel: +33 4 73 75 06 04, Fax: +33 4 73 75 06 19, e-mail: a_lelong@chu-clermontferrand.fr

Background: Since 2008, French health institutions providing medical, surgical and obstetrical care are assessed on the basis of a set of quality indicators. The French National Authority for Health developed a survey design in which 80 records are randomly selected from each institution. The main aim was to assess the effects of internal heterogeneity of a hospital that comprises several units. The survey method is based on the hypothesis of intra-institution homogeneity, which overlooks the fact that in wide hospitals homogeneity is related to departments and thus leads to overall intra-hospital heterogeneity. Methods: Simulated databases were created to modelise the heterogeneity of our hospital and computed to assess the reliance of indicator measurement. We used real data from a large teaching hospital having internal heterogeneity related to each department. Results: Variance under heterogeneity was greater than under homogeneity (3- to 18-fold) leading to an increased size of the confidence interval (CI) (at 95%) from 9 (given Haute Autorité de Santé sources) to 22 (for greatest internal heterogeneity). Conclusions: The variations in a quality indicator can be explained by intra-institution heterogeneity and are not related to changes in the quality policy of the hospitals and may lead to errors in terms of pay for performance.

Introduction

Since 2008, French health institutions providing medical, surgical and obstetrical care shall participate in the Indicators for Improvement of Care and Safety Quality (IPAQSS) survey, yearly. It is a retrospective survey that consisted of a collection of a set of data from which quality indicators are calculated by the French National Authority for Health (Haute Autorité de Santé—HAS), such as the patient record keeping. The hospital medical records of 80 patients are randomly selected from each institution based...
on the hypothesis of intra-structure homogeneity. The results are published, used for hospital accreditation and are planned to be used for adjusting the funding of hospitals by diagnosis-related groups on a pay-for-performance basis as in Anglo-Saxon countries.

Public authorities set the compliance threshold at 80%—from 0% if no HAS criteria are met to 100% if records are in full compliance with HAS criteria. In 2011, of the 1206 health institutions participating in the survey, 566 (47%) had results significantly lower than national targets. In the teaching hospital (CHU: Centre Hospitalier Universitaire) of Clermont-Ferrand, central France, the results of quality indicators vary widely from one year to another without ever reaching the 80% threshold. For example, the indicator for patient record keeping was 55% in 2010 and 69% in 2011. These variations occurred, despite the fact that there was no change in quality policy (the results of year n − 2 being published, while those of n − 1 were being collected). We hypothesized therefore that the variations in the indicator scores were due to significant intra-structure variability due to the large number of units having widely different results. In contrast, fewer variations were observed for individual results of each unit meaning that quality indicator measurements are more robust. It is noteworthy that intra-unit heterogeneity is less significant than inter-department heterogeneity. This is especially accurate for patient record keeping or pain management that closely reflect the quality policy of each unit.

The main aim of our study was to assess the effects of internal heterogeneity of a hospital that comprises several units.

Methods

Creation of tables for simulating the impact of heterogeneity

As individual assessment of patient record keeping was not accessible, our work was based only on the data of the IPAQSS surveys. Mean scores with their standard deviations (±SD) for each department and for the whole hospital were available. The effects of internal heterogeneity were evaluated by simulating different databases corresponding to one real situation as well as some theoretical ones.

Our study was based on simulated databases, containing 77,426 hospital stays (i.e. the recorded number of stays in the CHU of Clermont-Ferrand in 2011) and dividing between 99 units according to the annual number of hospital stays per unit. The impact of internal heterogeneity can be therefore modelled in our health-care institution on the basis of real data. Of the 99 units, 53 had been randomly selected at least once during the 4 years of the IPAQSS surveys (based on a hospital sample of 80 records a year). For these 53 departments, mean scores ± SD were known. For the remaining 46 units, the same proportion of records with a score greater than or equal to 80 and of records with a score less than 80 were applied, according to the results of the 53 units. Five sets of data were simulated:

Base 1—under the hypothesis of homogeneity (homogeneity base). It was made up of scores following a uniform distribution around the mean score of 80, the compliance threshold established by the HAS.

Base 2 and Base 3—three theoretical models under a hypothesis of heterogeneity. Base 2 with the greatest possible heterogeneity is made up of 50% of scores equal to 0 and 50% equal to 100 in the 77,426 patient records. Base 3 with moderate heterogeneity is composed of 25% of scores equal to 0, 25% equal to 100 and the remainder equal to 80 in the 77,426 patient records.

Base 4 and Base 5—under a hypothesis of realistic heterogeneity (i.e. the real model represented a base close to reality from real scores attributed to the CHU of Clermont-Ferrand over the period 2008–2011). Base 4 with a forcing of heterogeneity. The mean of the four scores obtained by each unit over the 4-year period has been calculated. If the mean was strictly lower than 80, the lowest score was recorded. Conversely, if the mean was greater than or equal to 80, the highest score was registered. Base 5 corresponded to a smoothed 4-year model. The global mean of the mean scores and the common variance registered. Base 5—under a hypothesis of realistic heterogeneity (i.e. the real model represented a base close to reality from real scores attributed to the CHU of Clermont-Ferrand over the period 2008–2011).

The mean of the four scores obtained by each unit over the 4-year period has been calculated. If the mean was strictly lower than 80, the lowest score was recorded. Conversely, if the mean was greater than or equal to 80, the highest score was registered. Base 5 corresponded to a smoothed 4-year model. The global mean of the mean scores and the common variance registered. Base 5—under a hypothesis of realistic heterogeneity (i.e. the real model represented a base close to reality from real scores attributed to the CHU of Clermont-Ferrand over the period 2008–2011).

Later, the mean of the four scores obtained by each unit over the 4-year period has been calculated. If the mean was strictly lower than 80, the lowest score was recorded. Conversely, if the mean was greater than or equal to 80, the highest score was registered. Base 5 corresponded to a smoothed 4-year model. The global mean of the mean scores and the common variance registered. Base 5—under a hypothesis of realistic heterogeneity (i.e. the real model represented a base close to reality from real scores attributed to the CHU of Clermont-Ferrand over the period 2008–2011).

The mean of the four scores obtained by each unit over the 4-year period has been calculated. If the mean was strictly lower than 80, the lowest score was recorded. Conversely, if the mean was greater than or equal to 80, the highest score was registered. Base 5 corresponded to a smoothed 4-year model. The global mean of the mean scores and the common variance registered. Base 5—under a hypothesis of realistic heterogeneity (i.e. the real model represented a base close to reality from real scores attributed to the CHU of Clermont-Ferrand over the period 2008–2011).

Drawing random samples

To follow the procedure of the IPAQSS survey devised by the HAS, a consistent selection of 80 records was made (i.e. using a random variable from a uniform distribution) from each of the databases created and a mean was calculated at each draw. Selection was performed k times (three values were tested: k = 100, k = 500 and k = 1000) for each of the bases. The variance of each sample composed of the mean scores collected k times was calculated with its 95% CI (figure 1).

Intra-class correlation coefficient

Intra-hospital heterogeneity in sampling can also be demonstrated by solving randomization of clusters. Each department can be assimilated to a cluster, taking into account the complexity of
Table 1 | Quality scores obtained for the indicator patient record keeping by the teaching hospital of Clermont-Ferrand between 2008 and 2011

<table>
<thead>
<tr>
<th>Year</th>
<th>Quality score</th>
<th>[95% CI]</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>68</td>
<td>[64–72]</td>
<td>18.26</td>
</tr>
<tr>
<td>2009</td>
<td>64</td>
<td>[60–69]</td>
<td>22.76</td>
</tr>
<tr>
<td>2010</td>
<td>55</td>
<td>[50–60]</td>
<td>21.83</td>
</tr>
<tr>
<td>2011</td>
<td>69</td>
<td>[64–74]</td>
<td>22.37</td>
</tr>
</tbody>
</table>

Table 2 | Results according to the number of random samples for the simulated databases

<table>
<thead>
<tr>
<th>k</th>
<th>Min–Max</th>
<th>[95% CI]</th>
<th>Width</th>
<th>Variance</th>
<th>Variance ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>35.00–65.00</td>
<td>[38.75–62.50]</td>
<td>22.50</td>
<td>33.50</td>
<td>18.21</td>
</tr>
<tr>
<td>500</td>
<td>35.00–66.25</td>
<td>[40.00–61.25]</td>
<td>21.25</td>
<td>30.36</td>
<td>17.55</td>
</tr>
<tr>
<td>1000</td>
<td>33.75–70.00</td>
<td>[37.55–82.60]</td>
<td>5.05</td>
<td>1.76</td>
<td>1</td>
</tr>
</tbody>
</table>

Simulated heterogeneity

Greatest heterogeneity

<table>
<thead>
<tr>
<th>k</th>
<th>Min–Max</th>
<th>[95% CI]</th>
<th>Width</th>
<th>Variance</th>
<th>Variance ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>52.75–74.00</td>
<td>[56.25–71.75]</td>
<td>15.5</td>
<td>16.85</td>
<td>9.16</td>
</tr>
<tr>
<td>500</td>
<td>52.75–77.75</td>
<td>[56.25–72.00]</td>
<td>15.75</td>
<td>17.82</td>
<td>10.30</td>
</tr>
<tr>
<td>1000</td>
<td>50.75–77.75</td>
<td>[56.00–73.00]</td>
<td>17.00</td>
<td>18.51</td>
<td>10.52</td>
</tr>
</tbody>
</table>

Simulated heterogeneity

<table>
<thead>
<tr>
<th>k</th>
<th>Min–Max</th>
<th>[95% CI]</th>
<th>Width</th>
<th>Variance</th>
<th>Variance ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>43.70–61.47</td>
<td>[47.21–59.46]</td>
<td>12.25</td>
<td>10.54</td>
<td>5.73</td>
</tr>
<tr>
<td>500</td>
<td>43.70–61.47</td>
<td>[47.92–59.46]</td>
<td>11.54</td>
<td>9.35</td>
<td>5.40</td>
</tr>
<tr>
<td>1000</td>
<td>43.70–62.80</td>
<td>[47.93–59.46]</td>
<td>11.53</td>
<td>8.92</td>
<td>5.07</td>
</tr>
</tbody>
</table>

Smoothed 4-year model

<table>
<thead>
<tr>
<th>k</th>
<th>Min–Max</th>
<th>[95% CI]</th>
<th>Width</th>
<th>Variance</th>
<th>Variance ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>59.36–70.61</td>
<td>[60.11–69.27]</td>
<td>9.16</td>
<td>6.44</td>
<td>3.50</td>
</tr>
<tr>
<td>500</td>
<td>57.72–70.63</td>
<td>[59.72–69.26]</td>
<td>9.54</td>
<td>6.14</td>
<td>3.54</td>
</tr>
<tr>
<td>1000</td>
<td>57.18–71.75</td>
<td>[60.01–69.22]</td>
<td>9.21</td>
<td>4.35</td>
<td>2.47</td>
</tr>
</tbody>
</table>

a: Width of the CI.
b: The ratio is defined as follows: \( \sigma^2 \) (base \( X \))/\( \sigma^2 \) (base homogeneity).

Table 3 | Calculation of the ICC, design effect (\( D_{eff} \)) and adjusted number of records for each base

<table>
<thead>
<tr>
<th>Hypotheses</th>
<th>ICC</th>
<th>( D_{eff} )</th>
<th>Adjusted number of records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greatest heterogeneity</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Moderate heterogeneity</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Homogeneity</td>
<td>( \sim 10^{-3} )</td>
<td>( \sim 1 )</td>
<td>80</td>
</tr>
<tr>
<td>Forced homogeneity</td>
<td>0.67</td>
<td>53.99</td>
<td>4319</td>
</tr>
<tr>
<td>Smoothed 4-year model</td>
<td>0.42</td>
<td>34.43</td>
<td>2754</td>
</tr>
</tbody>
</table>

Materials

The simulated databases were created with SAS software version 9.3 (SAS Institute, Inc., Cary, NC).

Results

Population size was the same for the five sets of data simulated (\( N = 77\,426 \)) and corresponded to the number of patients hospitalized in the CHU of Clermont-Ferrand in 2011. Base 1 (homogeneity base) had a mean score of 80 (SD ± 11.51). For databases based on the hypothesis of heterogeneity (base 2 and base 3), the mean ranged from 50 (SD ± 50) to 65 (SD ± 38.41). For bases based on the hypothesis of realistic heterogeneity, the mean was 54 (SD ± 26.34) for base 4 (base with forcing of heterogeneity) and 65 (SD ± 20.92) for base 5 (smoothed 4-year model) (Supplementary table). The results of our simulations were comparable with those recorded during the national indicator assessment survey. The actual mean scores of the CHU in Clermont-Ferrand ranged from 55 to 69 (table 1).

For databases based on the hypothesis of heterogeneity (base 2 and base 3), the variance and width of the CI rose in line with the increase in heterogeneity. The width of CI ranged from 5 (base 1: homogeneity base) to 22 (base 2: model of greatest possible heterogeneity) with a corresponding fluctuation in variance from 1.76 to 33.50. For bases based on the hypothesis of realistic heterogeneity, the CI varied from 9 (base 5: smoothed 4-year model) to 12 (base 4: with forcing of heterogeneity) with a corresponding range in variance from 4.35 to 10.54. Wide variations were observed when the variance of bases 2 to 5 was compared with base 1 (homogeneity base). The variance of base 2 (base of greatest possible heterogeneity) is 18-fold greater than that of base 1 (homogeneity base), whereas the variance of base 3 (moderate heterogeneity) is about 10 times greater. If a comparison is made with bases based on the hypothesis of heterogeneity, the variance is about 3–6-fold than that of base 1 (homogeneity base) (table 2).

Using the correction of sampling by ICC, setting the number of records to be assessed at 80 would be valid based on the hypothesis of homogeneity. In contrast, based on the hypothesis of heterogeneity, the estimated number would be between 2754 and 4319 (table 3).

Discussion

Main results

The results of the IPAQSS surveys were not robust owing to the internal heterogeneity of hospitals. In the models of inhomogeneity, variance was greater, which is responsible for highly fluctuated results. This variability can mainly explain the differences in overall scores from one year to another, by producing effects that go far beyond changes in indicators related to hospital policy.

Comparisons with other studies

These quality indicators were developed by the coordination for performance measurement and improvement of hospital quality project (COMPAQH: COrdonnatiion pour la Mesure de la Performance et l’Amélioration de la Qualité Hospitalière). The relevance and metrological quality of these indicators have been tested and validated by the developers. The COMPAQH study showed that samples of 60 to 100 cases are sufficient to express differences between health institutions. Their conclusions are based on two studies. The first one established the appropriate sample size to obtain a fixed CI for a proportion (i.e., a sample size of 100 would be sufficient to have a CI of 95% and standard deviation 0.2). However, the main judgement criterion for the indicator of patient record keeping is not a proportion but a mean. If the formula of the number of subjects required for means to obtain a fixed-width CI with SD 0.2 is applied, then a...
sample size of 128,090 is required. According to the second study, a sample size of 60 individuals would be sufficient to observe a difference. However, these authors were comparing two health plans, one involving 505 subjects and the other 253, using a stratified random sampling method with low within-stratum variances. This type of estimate is not applicable to a hospital, which has large internal heterogeneity. Teaching hospitals in France can admit up to and beyond 100,000 patients annually; there were 77,426 hospital stays in 2011 in the CHU of Clermont-Ferrand, while in the same year the figure in the CHU of Bordeaux reached 179,714 (combined total of stays of more and less than 1 day). In contrast, the number of departments varies over 60 to 150 and heterogeneity between them is the same in each hospital.

Moreover, the association between procedure and outcomes quality indicators (measured in France and including the patient record keeping) should be established. Romano et al. compared procedure and outcomes quality indicators with the Patient Safety indicators (PSI) as performance indicators. Because of moderate sensitivity and positive predictive values, they concluded that further validation should be considered before comparing or rewarding hospital performance. In France, just one part of the PSI is currently collected: the indicator of surgical site infection (ISO: Infections du Site Opératoire). In a previous work (unpublished), we calculated the correlation coefficient (r of Pearson) between the results of the ISO and the French patient record keeping, for 95 universities and regional hospitals practicing surgery. There was no correlation between these two quality indicators (r = −0.005, not significant). Another French study by Dang et al. showed no relationship between the quality of patient record keeping (the level of traceability of pain) and patient’s satisfaction with pain control.

Implications
The assessment of quality indicators for wide health institutions with large internal heterogeneity and structured in clusters (units in hospital setting) can be thwarted by a bad survey design. The internal heterogeneity of health institutions related to inter-department heterogeneity and so to the number and size of unit per hospital should be considered in the protocol of the survey. Then, it would be possible to adapt the survey design as for cluster surveys taking into account the drawback, by assessing a greater number of medical records from 2754 to 4319, which is 34 to 54 times greater than that based on the hypothesis of homogeneity; with such size of sample, all the units are correctly represented in every sampling process.

Because most medical records are computerized, another way would be to use virtually exhaustive analyses of computerized databases. However, such a survey would require data to be managed via distributed computing, an option that is not universally available.

Strength and limits
The data available were only the mean scores with their SD for the 53 departments randomly selected from the IPAQSS surveys. Consequently, to model results for health institutions like our, choices were made to consider the same proportion of units which had scores greater than or equal to 80 to be as close as possible to the reality.

Our study showed the actual problem the IPAQSS survey faced. To consider this problem, an estimation of the numbers of records needed to have a reliable indicator should be measured. We estimated between 2754 and 4319 the numbers of medical records that should be included, this is probably not possible considering human resources. Another solution should be the implementation of distributed computing system. Nonetheless, further statistical strategies are needed to find other concrete solutions to solve the problem and how these solutions should be implemented.

Conclusion
Our study showed that the results of the IPAQSS surveys were not robust depending on the internal heterogeneity of health institutions. The hypothesis of homogeneity does not correspond to the real-life situation of wide hospitals. Thus, any financial incentives and/or modulations based on such indicators should take into account the level of internal heterogeneity and/or use of distributed computing to obtain assessments as close as possible to reality.

Supplementary data
Supplementary data are available at EURPUB online.

Acknowledgements
Our work has been presented orally and by poster previously at the 6th European Public Health Conference at Brussels in November 2013. Our work has also been presented orally previously at the ADELF-EMOIS Congress at Paris in April 2014.

Funding
This research was supported by our university hospital and by a grant from the French National Agency for Research (ANR-10-TECSAN-008-01/GINSENG project).

Conflicts of interest: None declared.

Key points
- The survey method by drawing lots of 80 records based on intra-structure homogeneity hypothesis to develop quality indicators is not suitable in wide hospitals because it neglects homogeneity related to units and thus the global intra-hospital heterogeneity.
- All incitation and/or financing modulation based on such indicators should take into account the level of intra-structure heterogeneity and/or use distributed computerization to obtain measures close to reality.
- Assessment of formal data in the French hospitals should be based on the implementation of distributed computing system.

References
4 Haute Autorité de Santé (HAS). Available at: www.has-sante.fr (5 February 2015, date last accessed).
Introduction

Life expectancy (LE) has been increasing in most European countries and further increases have been predicted. Rising LE accelerates the growth of the elderly population which creates an additional burden on the health care and pension systems. To counter this burden, almost all western countries are seeking ways to bring retirement ages more in line with increases in LE. The success of such a policy depends crucially on to what extent the additional life years that make up the increased LE are spent in good health.

The empirical literature suggests that poor health reduces the capability to work which decreases labour force participation. Popular measures of health used in studies linking health to labour participation include self-rated health and disability. Most studies focused on the elderly, and the results suggest that health is the most important determinant of labour participation for older workers. Although many studies found a negative effect of health losses on labour participation, there is lack of consensus on the strength of the effect. Estimates of the causal effect of health on labour participation are sensitive to the choice of health measure, identification assumptions regarding the effect of health on labour status and on the institutional context, such as disability insurance schemes or early retirement policies.

Measures of health such as disability status can be used to calculate disability-free life expectancy (DFLE). DFLE is a measure of population health that might be more relevant than LE for the political debates regarding the ageing of the population as it captures not only information on length of life but also on how many years are spent in good health. Even though the strength of the estimated impact of health on labour market status varies between studies, all studies suggest that productivity and labour participation can be extended by preventing or postponing health problems. Furthermore, no matter how disability is defined, disabled people seem to have on average lower levels of employment and income than non-disabled people. Disabled people also have lower working life expectancies (WLEs).

This study was designed to answer a question not before directly addressed in the literature on the relation between health and employment: does living longer in good health lead to a longer working life? This question is important from a public health and a social economic perspective, considering the consequences for political decision making regarding the pension system. Specifically, we aim to answer the question to what extent a decline in disability incidence could influence DFLE, and from there WLE.