Effect of data validation audit on hospital mortality ranking and pay for performance

Skerdi Haviari,[1,2] François Chollet,[1] Stéphanie Polazzi,[1] Cecile Payet,[1] Adrien Beauveil,[1] Cyrille Colin,[1,2] Antoine Duclos[1,2]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bmjqs-2018-008039).

[1]Pôle Information Médicale Evaluation Recherche, Hospices Civils de Lyon, Lyon, France
[2]Health Services and Performance Research lab (HESPER EA7425), Université Claude Bernard Lyon 1, Lyon, France

Correspondence to
Skerdi Haviari, Pôle Information Médicale Evaluation Recherche, Hospices Civils de Lyon, Lyon 69003, France;
skerdihaviari@gmail.com

Received 2 March 2018
Revised 27 June 2018
Accepted 5 October 2018

To cite: Haviari S, Chollet F, Polazzi S, *et al*. *BMJ Qual Saf* Epub ahead of print: [please include Day Month Year]. doi:10.1136/bmjqs-2018-008039

ABSTRACT

Background Quality improvement and epidemiology studies often rely on database codes to measure performance or impact of adjusted risk factors, but how validity issues can bias those estimates is seldom quantified.

Objectives To evaluate whether and how much interhospital administrative coding variations influence a typical performance measure (adjusted mortality) and potential incentives based on it.

Design National cross-sectional study comparing hospital mortality ranking and simulated pay-for-performance incentives before/after recoding discharge abstracts using medical records.

Setting Twenty-four public and private hospitals located in France

Participants All inpatient stays from the 78 deadliest diagnosis-related groups over 1 year.

Interventions Elixhauser and Charlson comorbidities were derived, and mortality ratios were computed for each hospital. Thirty random stays per hospital were then recoded by two central reviewers and used in a Bayesian hierarchical model to estimate hospital-specific and comorbidity-specific predictive values. Simulations then estimated shifts in adjusted mortality and proportion of incentives that would be unfairly distributed by a typical pay-for-performance programme in this situation.

Main outcome measures Positive and negative predictive values of routine coding of comorbidities in hospital databases, variations in hospitals' mortality league table and proportion of unfair incentives.

Results A total of 70 402 hospital discharge abstracts were analysed, of which 715 were recoded from full medical records. Hospital comorbidity-level positive predictive values ranged from 64.4% to 96.4% and negative ones from 88.0% to 99.9%. Using Elixhauser comorbidities for adjustment, 70.3% of hospitals changed position in the mortality league table after correction, which added up to a mean 6.5% (SD 3.6) of a total pay-for-performance budget being allocated to the wrong hospitals. Using Charlson, 61.5% of hospitals changed position, with 7.3% (SD 4.0) budget misallocation.

Conclusions Variations in administrative data coding can bias mortality comparisons and budget allocation across hospitals. Such heterogeneity in data validity may be corrected using a centralised coding strategy from a random sample of observations.

INTRODUCTION

Hospital administrative databases are a readily available source of data that can be used for pay for performance. One of the outcomes of interest is mortality due to its lack of ambiguity. In the context of quality of care, mortality is often measured and compared at the hospital level, and the hospitals with lowest mortality can be rewarded with financial incentives.[1–3] For example, this has been done, with mixed results, in two large incentivisation programmes: the Hospital Quality Incentive Demonstration (HQID; USA, 2003) and Advancing Quality (AQ; UK, 2008).[1 3 4]

One of the main challenges when doing these comparisons is that, in order to make fair judgements, one needs to account for the fact that some hospitals admit sicker patients than others. This is often done using administrative insurance claims data, which records the patient health status using standardised codes. These codes, for many different diseases, can be combined into reduced lists with broader categories that are most meaningful to adjust for risk of death at admission. Two widely used such lists are the Charlson and Elixhauser indexes.[5 6] The Charlson index (17 diseases) was initially derived from medical records[7] and was later adapted to hospital administrative data,[8] whereas the Elixhauser index (31 diseases) was originally designed to be used with administrative data.[9]

The assumption that data expressed with such standardised codes are valid, meaningful and actually standardised is not necessarily true. For example, some conditions may be associated with higher payouts, because sicker patients require more care, and some hospitals may want to game the payout system, leading them to exaggerate pre-existing diseases in their patients. Conversely, some comorbidities that are not associated with higher payouts may not be reported by some hospitals with overstretched coding staff. Even before the

coding step, some diseases can be easier or harder to diagnose clinically in different regions and populations. These issues belong to the umbrella term 'constant risk fallacy', meaning that the same coded variable can have different validity or mean different things in different settings and may therefore not be associated with a single, constant risk. This could make adjusted comparisons meaningless.[10–12]

This is very problematic in the context of pay-for-performance programmes, because how well hospitals and physicians respond to incentives is linked to how fair the evaluation method is.[13] To date, the potential impact on fairness of these variations in database codes' validity, which we call heterogeneous validity, has rarely been quantified. Single-centre and/or single-disease comparisons using medical records or registers have been published,[14 15] but multicentre all-patient comparisons relevant to overall performance evaluation are still needed. Here, we designed a new analytical strategy using partial centralised recoding and used it to evaluate the influence of interhospital variations in disease coding validity on hospital mortality rankings and potential performance incentives.

METHODS

The study design is summarised in figure 1. Details can be found in the technical appendix and in references.[16 17]

Population and data sources

Fifty diverse acute care hospitals were invited to participate (twenty-four accepted). Electronic discharge abstracts were retrieved for all inpatients from 2010 belonging to one of the 78 most deadly diagnosis-related groups (a type of administrative classification based on burden of care). Thirty abstracts were randomly drawn and centrally recoded from full electronic medical charts by two experienced coding

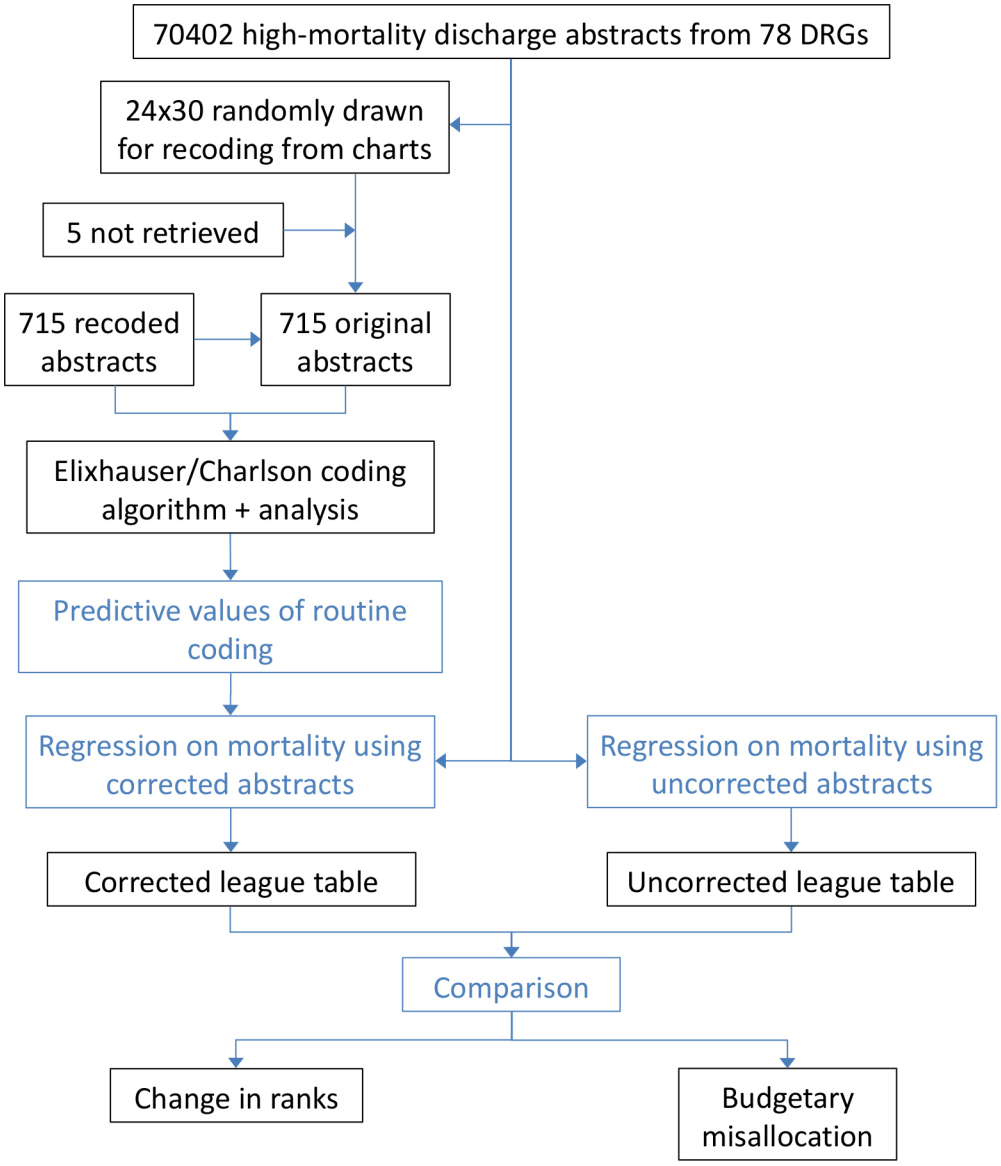

Figure 1 Study flow chart. DRGs, diagnosis-related group.

professionals including one physician, who were instructed to code for maximum epidemiological accuracy rather than financial optimisation. Differences were resolved by consensus. This provided a reference standard against which to check the validity of routine coding in a selected sample and then extrapolate to the whole population which, in turn eliminated interhospital coding variability. This reference standard is not a gold standard in an absolute sense, because the central recoders may have their own biases and habits.

Uncorrected hospital mortality

We ran a classical multivariate analysis using the Charlson or Elixhauser comorbidities indices from administrative billing codes as covariates, in addition to patient age, sex, area earnings, emergency admission and discharge month.

Each hospital was also a covariate; this allowed a quantification of the extent to which each hospital was, figuratively, a protective or risk factor for death and in turn a ranking of the hospitals from best to worst based on this criterion.

Validity of comorbidity codes

The previous multivariate analysis takes administrative comorbidity codes at face value, but we also wanted to correct for errors in those codes. For this, as we randomly drew 30 charts from each hospital and recoded them centrally. This provided a common reference standard, to which the initial coding done in each hospital was compared. We used this sample to derive a probability, for each hospital/comorbidity pair, that the comorbidity was actually there when the hospital said so (positive predictive value (PPV)) and, conversely, that the comorbidity was absent if the hospital did not mention it (negative predictive value (NPV)).

To this end, a Bayesian hierarchical model was run. The intuition is the following (see appendix for details): compared with the average, the model considers each hospital to be somewhat better or worse at coding, and each comorbidity to be somewhat easier or more difficult to code for properly, based on the randomly recoded charts. On top of this, the model adds a small improvement or deterioration if a particular hospital seems to be particularly good or bad at coding a particular disease. By integrating these three components for each hospital/comorbidity pair, specific PPVs and NPVs are obtained. Because the sample size is limited, these PPV and NPV are uncertain, with a probability distribution that represent how valid the codes are likely to be.

Corrected hospital mortality

When doing the first multivariate analysis of mortality, we took the hospital comorbidity codes at face value: for example, when heart failure was coded, the model added the full extra risk of death associated with heart failure.

Using the PPV/NPV, we could correct for the tendency of hospitals to exaggerate (overcode) or understate (undercode) using a corrected version of the initial mortality model. In the corrected version, if a particular hospital tended to exaggerate, for example, if there actually was heart failure only half the time that hospital coded it, the corrected model added only half the risk associated with heart failure for patients from that exaggerating hospital. The same logic was applied to all comorbidites.

This allowed us to compute a new risk of death associated with each hospital, both adjusted on the comorbidities and corrected for the tendency of hospitals to misstate them. Because validity (PPV/NPV) estimates were probabilistic, we repeated the corrected computation 1000 times, resulting in a Monte Carlo analysis. For this, we drew 1000 sets of PPV/NPV values from the Bayesian posterior.

Pay for performance

Then, variations in hospital mortality were applied to a tournament pay-for-performance system modelled after AQ and HQID, in which hospitals with the lowest adjusted mortality are ranked higher and better paid, resulting in 1000 corrected rankings and payouts. We then computed the difference in payouts between the corrected and uncorrected versions and estimated a possible size of unfair payouts due to hospital misstatements of comorbidities, expressed as percentage of the total incentive payouts. The 1000 repeats allowed us to provide a reasonable average and quantify uncertainty around our estimate.

Sensitivity analyses

We ran several sensitivity analyses: (1) to check whether code validity was similar in different patient populations, (2) whether more refined pay-for-performance schemes were more vulnerable to heterogeneity, (3) whether different Bayesian models were more robust and (4) to what extent additional recoded data would have been useful.

RESULTS

Among the 720 medical records randomly selected from 24 participating hospitals, 5 could not be retrieved for recoding discharge abstracts and were excluded from analysis (figure 1). Table 1 shows the side-by-side distribution of the 715 discharge abstracts before and after recoding and the 70 402 inpatient stays from which they were randomly selected. The validity of the Elixhauser and Charlson comorbidities across all hospitals is shown in table 2. For the Elixhauser index, the median value of estimated hospital-level coding PPV ranged from 70.0% for weight loss (IQR for hospitals, 51.4–80.8) to 96.4% (94.8–97.5) for cardiac arrhythmia, while estimated negative

ones ranged from 89.0% for uncomplicated hypertension (81.1–92.2) to 99.9% for AIDS/HIV (99.9–99.9). For the Charlson index, estimated positive values ranged from 64.4% (57.4–69.5) for liver disease to 90.3% for paraplegia or haemiplegia (87.4–92.0), while estimated negative ones ranged from 94.8% for localised malignancies (92.2–95.3) to 99.9% for AIDS/HIV (99.9–99.9).

Figure 2 presents changes in hospitals' position after correction for either Elixhauser or Charlson comorbidities, compared with the average mortality on funnel plots (figure 2A, B). These changes were more explicit when depicting hospitals positions against each other on league tables (figure 2C, D). Simulated AQ and HQID programmes found that hospitals had a 70.3% and 61.5% probability to experience change in ranking position after correction of either Elixhauser (figure 2E) or Charlson (figure 2F) indexes, respectively. For the Elixhauser index, a drop in the league table, of a mean 2.2 ranks, was computed for 35.1% of hospitals, whereas a gain of a mean 2.2 ranks was computed for 35.1% of hospitals. For the Charlson index, there was a mean drop of 1.8 ranks for 30.1% hospitals and a mean gain of 1.8 ranks for 31.4% hospitals. These shifts translated into coding-related undeserved budget bonuses of 2% for 4.9% of hospitals and penalties of 2% for 4.9% of hospitals with the Elixhauser index; the remainder had unchanged budgets (figure 2G). For the Charlson index, 5.5% of hospitals had a similar undeserved 2% budget increase/decrease, with exceptionally unlikely (<0.05%) cases of 4% loss (figure 2H). At the programme level, the proportion of all improperly allocated bonuses and penalties (not accounting for hospital size) reached a mean 6.5% (SD 3.6) of the total programme budget with the Elixhauser comorbidities adjustment and 7.3% (SD 4.0) with the Charlson comorbidities adjustment (figure 2I, J).

Online supplementary figures 1 and 2 present the results of sensitivity analyses, with higher improper allocation when incentives were distributed more smoothly across the league table (11.5%, SD 3.0%, for Elixhauser, and 10.5%, SD 3.1%, for Charlson).

Online supplementary figure 3 presents sensitivity analyses on the Bayesian network structure and recoding data. Removing the interaction terms led to a different distribution of errors and a lower uncertainty around the estimate. A triplicated data model, artificially triplicating each observation to a total of 90 charts per hospital, led to more precise estimates and a reduction in the probability of zero misallocation.

DISCUSSION

The advent of information technology has provided an opportunity to collect and analyse data on a large scale to move closer to value-based care, but the quality of the data and its impact on the relevance of the analyses often go unquestioned. Heterogeneous coding, in particular, has been singled out as a particularly difficult issue, and differences between genuinely differential mortality and artefacts due to heterogeneous coding have puzzled researchers, which led Mohammed *et al*, in an often cited study,[12] to claim that 'unfortunately, no statistical method exists for teasing [them] apart'.

Here, in the particular context of analysing hospital mortality, we have devised a method to estimate the effect of heterogeneous coding apart from that of actual differential mortality by using a central review to standardise coding. Using this approach, we show that heterogeneous coding practices are very likely to impact league tables based on adjusted mortality inferred from administrative data. The resulting error in distribution of incentives, for a tier-based programme, is around 7%. For AQ and HQID, total incentivisation budgets have reached £5 million (over the first 18 months)[4] and US$12 million USD (yearly),[18] respectively. This would translate into a substantial amount of resources being distributed either at random or to hospitals that are able to game the system. Given the sums at play, and depending on political resistance to perceived unfairness, one possibility for health policy makers is to implement the method we report here to correct for heterogeneous coding practices, provided an accepted gold standard can be agreed on. This could require significant manpower, since many medical records need to be re-examined for the correction to be precise, but it should be worth the effort in many cases.

The question of selection of medical records for centralised recoding is not trivial for several reasons. First, patients with few ailments will not allow for a good estimation of PPV due the lack of cases, thereby greatly reducing the effectiveness of the correction. Second, extrapolating validity is difficult, as suggested by high meta-analytical heterogeneity in validity studies,[19] by the fact that PPV and NPV vary with prevalence and, more importantly, by the fact that even sensitivity and specificity seem to vary greatly depending on the population. We therefore expect that it would be invalid to derive sensitivity and specificity values on a high-mortality set and then use them on a more general population, as is evident in our sensitivity analyses on the overall patient population.

Comparisons involving different case mixes are often not stratified properly in the literature,[20] but our results suggest as even if they were, having taken administrative codes at face value may have flawed their results. In particular, the lack of national selection criteria that are actually standardised in practice (and not only in theory) may produce substantially different results for calculating hospital-wide mortality.[21] Overall, we would argue either against the use of administrative data for case-mix adjustment in hospital performance benchmarking, or its restriction to high-prevalence populations, where some type of centralised coding heterogeneity check would be feasible. Furthermore,

Table 1 Characteristics of the discharge abstracts used for calculating mortality rates and of the random sample used for evaluation of estimating positive and negative predictive values

	78 DRGs, stays >0 night	Discharge abstracts before recoding	Discharge abstracts after recoding
	(N=70 402)	(N=715)	(N=715)
Quantitative variables, median (IQR)			
Age (years)	74 (60–83)	72 (61–82)	72 (61–82)
Length of stay (days)	8.0 (4.0–14.0)	8.0 (3.5–15.0)	8.0 (3.5–15.0)
Logical variables, n (%)			
Death during stay	7941 (11.3)	84 (11.7)	84 (11.7)
Female	31 632 (44.9)	322 (45)	322 (45)
Elixhauser comorbidities, n (%)			
Hypertension, uncomplicated	19 988 (28.4)	193 (27.0)	260 (36.4)
Cardiac arrhythmia	14 711 (20.9)	169 (23.6)	210 (29.4)
Congestive heart failure	11 304 (16.1)	115 (16.1)	139 (19.4)
Solid tumour without metastasis	8302 (11.8)	101 (14.1)	132 (18.5)
Weight loss	8130 (11.5)	82 (11.5)	67 (9.4)
Chronic pulmonary disease	7055 (10.0)	73 (10.2)	79 (11.0)
Renal failure	6908 (9.8)	62 (8.7)	72 (10.1)
Diabetes, uncomplicated	6579 (9.3)	75 (10.5)	85 (11.9)
Metastatic cancer	5984 (8.5)	77 (10.8)	92 (12.9)
Fluid/electrolyte disorders	5384 (7.6)	55 (7.7)	97 (13.6)
Diabetes, complicated	5088 (7.2)	45 (6.3)	47 (6.6)
Other neurological disorders	5080 (7.2)	42 (5.9)	56 (7.8)
Alcohol abuse	4558 (6.5)	42 (5.9)	41 (5.7)
Valvular disease	4207 (6.0)	41 (5.7)	59 (8.3)
Paralysis	4066 (5.8)	35 (4.9)	57 (8.0)
Obesity	3926 (5.6)	35 (4.9)	45 (6.3)
Peripheral vascular disease	3859 (5.5)	32 (4.5)	37 (5.2)
Depression	3761 (5.3)	28 (3.9)	28 (3.9)
Liver disease	3688 (5.2)	39 (5.5)	47 (6.6)
Hypertension, complicated	2328 (3.3)	24 (3.4)	27 (3.8)
Hypothyroidism	2235 (3.2)	23 (3.2)	29 (4.1)
Coagulopathy	1956 (2.8)	18 (2.5)	23 (3.2)
Deficiency anaemia	1760 (2.5)	16 (2.2)	22 (3.1)
Pulmonary circulation disorder	1728 (2.5)	12 (1.7)	24 (3.4)
Lymphoma	948 (1.3)	6 (0.8)	12 (1.7)
Rheumatoid arthritis	891 (1.3)	5 (0.7)	5 (0.7)
Blood loss anaemia	635 (0.9)	8 (1.1)	6 (0.8)
Psychoses	543 (0.8)	4 (0.6)	4 (0.6)
Drug abuse	320 (0.5)	2 (0.3)	1 (0.1)
Peptic ulcer disease	244 (0.3)	0 (0.0)	1 (0.1)
AIDS/HIV	147 (0.2)	2 (0.3)	2 (0.3)
Charlson comorbidities, n (%)			
Congestive heart failure	11 304 (16.1)	115 (16.1)	139 (19.4)
Chronic pulmonary disease	7055 (10.0)	73 (10.2)	79 (11.0)
Diabetes, uncomplicated	6916 (9.8)	83 (11.6)	96 (13.4)
Renal disease	6915 (9.8)	62 (8.7)	72 (10.1)
Cancer	6661 (9.5)	74 (10.3)	91 (12.7)
Metastatic cancer	5984 (8.5)	77 (10.8)	92 (12.9)
Cerebrovascular disease	5420 (7.7)	35 (4.9)	48 (6.7)
Dementia	4627 (6.6)	45 (6.3)	38 (5.3)
Diabetes, complicated	4175 (5.9)	31 (4.3)	34 (4.8)
Paraplegia and haemiplegia	4066 (5.8)	35 (4.9)	57 (8.0)

Continued

Table 1 Continued

	78 DRGs, stays >0 night	Discharge abstracts before recoding	Discharge abstracts after recoding
	(N=70 402)	(N=715)	(N=715)
Peripheral vascular disease	3859 (5.5)	32 (4.5)	37 (5.2)
Myocardial infarction	2124 (3.0)	19 (2.7)	31 (4.3)
Mild liver disease	1890 (2.7)	19 (2.7)	28 (3.9)
Moderate/severe liver disease	1553 (2.2)	16 (2.2)	12 (1.7)
Connective/rheumatic disease	711 (1.0)	3 (0.4)	4 (0.6)
Peptic ulcer disease	574 (0.8)	3 (0.4)	4 (0.6)
AIDS/HIV	147 (0.2)	2 (0.3)	2 (0.3)

Rounded individual ages are due to confidentiality restrictions. No data were missing for demographic and administrative variables. For data-derived comorbidities, the study design treats missing data as false negatives. All hospitals had 30 recoded abstracts regardless of their size, which causes some small discrepancies between the background set and the recoded set.
*DRG, diagnosis-related group.

because cross-sectional comparison of institutions is not ideally suited to monitor hospital performance prospectively, tracking mortality over time at the individual hospital level may be more efficient to reduce patient harm.[22] Assuming that the risk of misinterpretation due to variability in data coding and patient case-mix within the same hospital may be negligible compared with what is expected from one institution to another, each hospital would then be considered as its own performance benchmark.[23]

Keeping in mind that it was assessed by central recoders with their own potential biases, the overall quality of coding in our study was moderate, with significant variations between comorbidities (online supplementary table). Other studies for various specific indications such as venous thromboembolism,[24 25] arrhythmia,[26] stroke,[27] hypersensitivity reactions,[28] bone metastases,[29] glaucoma,[30] hip fractures,[31] hidradenitis suppurativa,[32] acute kidney injury,[33] overdoses[34] and sepsis[19 35] have also found average validity, with significant variations between studies for the same condition.[19 27 36] Under the French hospital payment system, patients with comorbidities are associated with higher payments to compensate for the higher burden of care, which may have favoured NPV over PPV and sensitivity over specificity. Conversely, the fact that most of the comorbidities are chronic conditions that are not necessarily a major component of the clinical picture seems to have had a strong effect in the opposite direction, as comorbidity rates in the corrected set mostly went up. The restriction to a high-mortality group increased prevalence to a certain extent, which may have favoured PPV at the expense of NPV compared with a more general population.

Our study is limited by its within-hospital sample size, which leads to uncertainty in corrected estimates of adjusted mortality and league tables based on them; our sensitivity analysis with triplicated data is an argument in favour of a higher sample size, around 90 per hospital, to have comfortably accurate estimates, especially if interaction terms have to be fitted. A weighed sampling strategy that avoids having variables with highly uncertain predictive value would also be useful. The same can be said of deaths which, in future iterations of the method, should also have a dedicated sampling strategy. Because of the design of incentives, statistical uncertainty may have added noise and increased the proportion of rewards counted as improperly allocated. However, our sensitivity analyses make it very unlikely that the main effect is due to noise. Another limitation is the validity of the reference standard itself which, to a certain extent, limits our conclusions to claims about heterogeneity rather than true validity. The limited number of participating hospitals introduces a potential response bias at the hospital level; a more general population might have more or homogeneous coding practices, and further work should test for this. The adjusted mortality model was not extensively refined and seems to have resulted in wider mortality variations than usually seen, but this is more likely to have reduced ranking shifts and misallocation estimates than increased them. We did not integrate other sources of error such as statistical inaccuracy of estimates or lack of granularity regarding administrative data to describe underlying patient risk, but these are expected to be similar in the two analyses with and without correction, so errors resulting from heterogeneous coding should be seen as distinct and added to them. Finally, we did not correct for heterogeneous validity of DRGs themselves, because this would have forced us to make imputations in the general, less severe hospital population, and we knew from sensitivity analyses that those would have been false; we are therefore likely to have underestimated the magnitude of bias due to heterogeneous validity.

Administrative coding improves with time, and it is possible that its quality could one day be sufficient to make comparisons between hospitals or healthcare providers that do not require correction. Our work has implications about how to measure and to improve data quality in healthcare but also more generally in all disciplines impacted by so-called big data, from

Table 2 Distribution of median estimates of PPV and NPV for hospital-comorbidity pairs

Median (IQR) (min-max), in %	PPV	NPV
Elixhauser comorbidities		
Hypertension, uncomplicated	95.8 (92.9–97.4) (70.9–98.6)	89.0 (81.1–92.2) (70.8–95.0)
Cardiac arrhythmia	96.4 (94.8–97.5) (72.5–98.8)	92.8 (90.0–95.5) (80.3–97.0)
Congestive heart failure	88.6 (75.7–92.5) (62.6–95.8)	94.8 (93.4–95.8) (85.2–97.1)
Solid tumour without metastasis	80.5 (72.8–87.3) (53.1–93.3)	93.6 (89.6–94.9) (74.1–95.5)
Weight loss	70.0 (51.4–80.8) (37.3–89.9)	97.6 (97.0–98.6) (91.4–98.9)
Chronic pulmonary disease	84.4 (77.5–89.2) (48.3–94.0)	97.1 (96.6–98.2) (92.3–98.7)
Renal failure	91.1 (86.6–93.6) (69.8–97.0)	98.0 (97.2–98.5) (94.1–98.9)
Diabetes, uncomplicated	86.3 (80.4–90.7) (53.1–95.5)	97.3 (95.9–97.9) (93.1–98.5)
Metastatic cancer	90.4 (88.1–93.9) (79.7–97.0)	96.9 (96.0–97.9) (90.6–98.5)
Fluid/electrolyte disorders	87.7 (83.2–93.5) (58.6–96.2)	93.7 (90.5–96.4) (86.8–97.5)
Diabetes, complicated	89.2 (83.5–92.3) (58.6–96.3)	98.9 (98.5–99.2) (97.7–99.4)
Other neurological disorders	93.3 (90.4–95.7) (77.1–97.9)	98.1 (97.6–98.5) (95.6–99.0)
Alcohol abuse	90.4 (83.1–93.3) (67.9–96.5)	99.4 (99.3–99.5) (98.9–99.7)
Valvular disease	90.0 (85.8–93.0) (60.7–96.5)	97.5 (96.4–98.3) (91.9–98.7)
Paralysis	93.3 (90.0–95.6) (76.8–97.7)	97.4 (96.5–97.8) (91.5–98.7)
Obesity	90.4 (86.6–93.6) (73.0–96.9)	98.4 (98.1–98.8) (96.2–99.2)
Peripheral vascular disease	84.8 (77.0–89.6) (53.6–94.4)	98.6 (98.1–98.9) (95.4–99.3)
Depression	88.1 (83.7–93.0) (72.7–96.4)	99.5 (99.4–99.6) (99.2–99.7)
Liver disease	86.2 (77.6–90.2) (63.4–95.3)	98.2 (97.4–98.6) (95.8–99.0)
Hypertension, complicated	92.0 (88.4–94.5) (73.3–97.3)	99.4 (99.3–99.6) (98.2–99.7)
Hypothyroidism	87.5 (83.5–92.1) (57.3–96.1)	99.0 (98.7–99.2) (97.0–99.4)
Coagulopathy	88.3 (83.7–92.1) (64.7–96.1)	99.2 (99.0–99.3) (97.6–99.5)
Deficiency anaemia	92.5 (88.2–95.0) (76.9–97.5)	99.2 (99.1–99.5) (98.1–99.6)
Pulmonary circulation disorder	93.6 (89.9–95.4) (76.8–97.7)	98.7 (98.4–99.0) (96.5–99.3)
Lymphoma	92.1 (88.5–94.6) (73.4–97.4)	99.4 (99.0–99.5) (98.8–99.7)
Rheumatoid arthritis	88.8 (83.4–92.3) (65.4–96.3)	99.8 (99.8–99.9) (99.7–99.9)
Blood loss anaemia	85.1 (76.7–89.8) (56.1–94.8)	99.8 (99.8–99.9) (99.6–99.9)
Psychoses	88.8 (83.4–92.1) (65.5–96.2)	99.8 (99.8–99.9) (99.7–99.9)
Drug abuse	87.1 (80.9–91.0) (61.7–95.5)	99.9 (99.9–99.9) (99.8–99.9)
Peptic ulcer disease	89.6 (84.6–92.8) (67.1–96.5)	99.8 (99.8–99.9) (99.7–99.9)
AIDS/HIV	90.6 (86.4–93.5) (69.7–96.8)	99.9 (99.9–99.9) (99.8–99.9)
Charlson comorbidities		
Congestive heart failure	85.7 (80.9–88.3) (76.0–90.8)	94.9 (92.3–95.3) (86.9–96.6)
Chronic pulmonary disease	80.6 (75.2–83.8) (67.0–87.0)	97.5 (96.4–97.9) (93.4–98.2)
Diabetes, uncomplicated	85.5 (81.3–88.4) (73.6–91.0)	96.8 (95.8–97.6) (92.8–97.9)
Renal disease	86.8 (84.1–89.9) (78.9–92.2)	97.7 (97.4–98.3) (95.3–98.4)
Cancer	67.5 (61.4–72.8) (54.1–78.5)	94.8 (92.2–95.3) (84.9–96.5)
Metastatic cancer	88.4 (85.1–90.2) (81.3–92.5)	97.3 (96.3–97.7) (90.8–98.0)
Cerebrovascular disease	71.3 (66.0–74.8) (57.5–81.6)	97.0 (96.0–97.8) (94.8–98.1)
Dementia	78.9 (72.9–82.9) (65.4–86.7)	99.5 (99.5–99.5) (99.0–99.6)
Diabetes, complicated	83.9 (80.2–86.4) (73.0–89.6)	99.1 (98.7–99.1) (98.0–99.2)
Paraplegia and haemiplegia	90.3 (87.4–92.0) (83.0–93.8)	97.2 (96.9–97.9) (93.2–98.2)
Peripheral vascular disease	78.6 (74.4–82.6) (66.8–86.2)	98.7 (98.2–98.8) (96.4–98.9)
Myocardial infarction	81.8 (77.2–85.1) (70.7–88.2)	98.4 (97.9–98.5) (95.9–98.7)
Mild liver disease	73.6 (67.5–77.8) (58.4–82.3)	98.1 (97.9–98.5) (96.1–98.7)
Moderate or severe liver disease	64.4 (57.4–69.5) (49.2–75.5)	99.4 (99.4–99.5) (99.1–99.5)
Connective tissue/rheumatic disease	84.9 (81.1–87.6) (75.1–90.5)	99.8 (99.8–99.8) (99.7–99.8)
Peptic ulcer disease	79.4 (74.3–82.8) (65.4–86.8)	99.7 (99.7–99.7) (99.4–99.8)
AIDS/HIV	83.6 (79.3–86.5) (73.0–89.5)	99.9 (99.9–99.9) (99.8–99.9)

The table is read as follows: for the coding of uncomplicated hypertension, the worst hospital had an estimated (median) PPV of 70.9%, the best of 98.6% and half of hospitals were in the 92.9 to 97.4 range.
NPV, negative predictive value; PPV, positive predictive value.

epidemiology to economics. Given the highly heterogeneous settings in which big data are collected, the saying 'garbage in, garbage out' is even more relevant for big data studies than for small-scale data analysis. Analysts should be wary of this weakness, and algorithms should be tuned in order to account for it. Scientists, policymakers and physicians should not disregard heterogeneous coding validity when they interpret findings and make judgements based on routinely collected data.

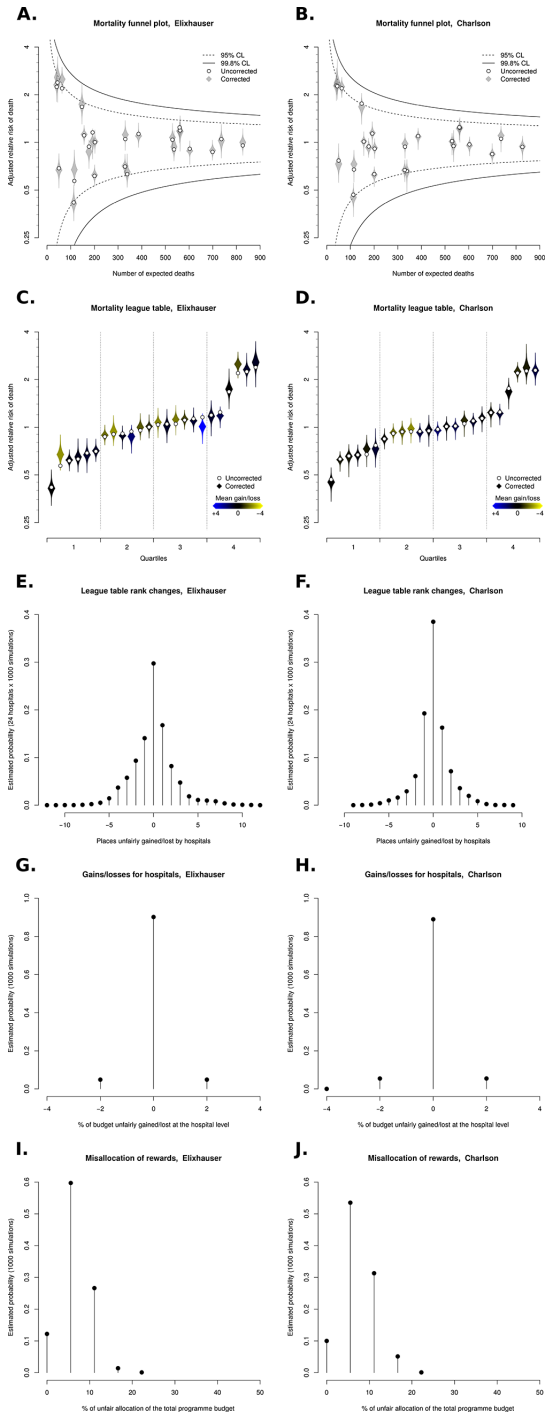

Figure 2 Shifts upon correction. The left side of the panel corresponds to an analysis with the Elixhauser comorbidity list (A, C, E, G and I); the right corresponds to the Charlson comorbidity list (B, D, F, H and J). (A and B) Funnel plots for adjusted relative risk of death. (C and D) Ranked forest plot of relative risk change upon correction, with mean gain/loss of ranks. Negative values are losses, whiles positive values are gains. The dispersion for corrected estimates is drawn with width proportional to percentile (the median has maximum width, quartiles have half the median width and so on). Regression coefficient uncertainty is not represented for either case. (E and F) Errors in hospital league table ranks due to heterogeneity. (G and H) Error in hospital-level financial bonuses. (I and J) Error in programme-level budget incentives allocation. Due to the granular structure of the payouts (which corresponds to AQ and HQID), only 19 values of misallocation are possible from 0 to 100%. AQ, Advancing Quality; HQID, Hospital Quality Incentive Demonstration.

Contributors AD and CC initiated and oversaw the project. SH and AD designed the analysis plan. AD, FC, SP, CP and AB obtained and pre-processed the data. SH, FC and SP analysed the data. SH and AD wrote the paper. All authors reviewed the paper.

Funding This study was funded by the French Ministry of Health.

Disclaimer The funding source had no involvement in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication. Researchers were independent from the funder.

Competing interests None declared.

Patient consent Not required.

Ethics approval This study was approved by the National Data Protection Commission (Commission Nationale de l'Informatique et des Libertés), in accordance with French ethical directives. We attest that we have obtained appropriate permissions and paid any required fees for use of copyright protected materials.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Original data can be shared with institutions that are authorised by the French Ministry of Health. Simulation results are available upon request.

REFERENCES

1 Flodgren G, Eccles MP, Shepperd S, *et al*. An overview of reviews evaluating the effectiveness of financial incentives in changing healthcare professional behaviours and patient outcomes. *Cochrane Database Syst Rev* 2011;7:CD009255.
2 Lee GM, Kleinman K, Soumerai SB, *et al*. Effect of nonpayment for preventable infections in U.S. hospitals. *N Engl J Med* 2012;367:1428–37.
3 Sutton M, Nikolova S, Boaden R, *et al*. Reduced mortality with hospital pay for performance in England. *N Engl J Med* 2012;367:1821–8.
4 McDonald R, Boaden R, Roland M, *et al*. A qualitative and quantitative evaluation of the Advancing Quality pay-for-performance programme in the NHS North West. *Health Services and Delivery Research* 2015;3:1–104.
5 Moore BJ, White S, Washington R, *et al*. Identifying increased risk of readmission and in-hospital mortality using hospital administrative data: the AHRQ elixhauser comorbidity index. *Med Care* 2017;55:698–705.
6 Southern DA, Quan H, Ghali WA. Comparison of the Elixhauser and Charlson/Deyo methods of comorbidity measurement in administrative data. *Med Care* 2004;42:355–60.
7 Charlson ME, Pompei P, Ales KL, *et al*. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* 1987;40:373–83.
8 Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J Clin Epidemiol* 1992;45:613–9.
9 Elixhauser A, Steiner C, Harris DR, *et al*. Comorbidity measures for use with administrative data. *Med Care* 1998;36:8–27.
10 van Gestel YR, Lemmens VE, Lingsma HF, *et al*. The hospital standardized mortality ratio fallacy: a narrative review. *Med Care* 2012;50:662–7.
11 Lilford R, Pronovost P. Using hospital mortality rates to judge hospital performance: a bad idea that just won't go away. *BMJ* 2010;340:c2016.

12 Mohammed MA, Deeks JJ, Girling A, *et al*. Evidence of methodological bias in hospital standardised mortality ratios: retrospective database study of English hospitals. *BMJ* 2009;338:b780.
13 Lilford R, Mohammed MA, Spiegelhalter D, *et al*. Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *Lancet* 2004;363:1147–54.
14 Shahian DM, Silverstein T, Lovett AF, *et al*. Comparison of clinical and administrative data sources for hospital coronary artery bypass graft surgery report cards. *Circulation* 2007;115:1518–27.
15 Stavem K, Hoel H, Skjaker SA, *et al*. Charlson comorbidity index derived from chart review or administrative data: agreement and prediction of mortality in intensive care patients. *Clin Epidemiol* 2017;9:311–20.
16 Leeflang MM, Deeks JJ, Rutjes AW, *et al*. Bivariate meta-analysis of predictive values of diagnostic tests can be an alternative to bivariate meta-analysis of sensitivity and specificity. *J Clin Epidemiol* 2012;65:1088–97.
17 Quan H, Sundararajan V, Halfon P, *et al*. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care* 2005;43:1130–9.
18 Centers for Medicare & Medicaid Services. Premier Hospital Quality Incentive Demonstration Fact Sheet, 2017. Available from: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/HospitalPremierPressRelease-FactSheet.pdf [Accessed 28 Sep 2017].
19 Jolley RJ, Sawka KJ, Yergens DW, *et al*. Validity of administrative data in recording sepsis: a systematic review. *Crit Care* 2015;19:139.
20 Payet C, Lifante JC, Carty MJ, *et al*. Methodological quality of surgical mortality studies using large hospital databases: a systematic review. *Ann Surg* 2017;265:1113–8.
21 Shahian DM, Wolf RE, Iezzoni LI, *et al*. Variability in the measurement of hospital-wide mortality rates. *N Engl J Med* 2010;363:2530–9.
22 Duclos A, Polazzi S, Lipsitz SR, *et al*. Temporal variation in surgical mortality within French hospitals. *Med Care* 2013;51:1085–93.
23 Bottle A, Jarman B, Aylin P. Strengths and weaknesses of hospital standardised mortality ratios. *BMJ* 2010;342:c7116.
24 Alotaibi GS, Wu C, Senthilselvan A, *et al*. The validity of ICD codes coupled with imaging procedure codes for identifying acute venous thromboembolism using administrative data. *Vasc Med* 2015;20:364–8.
25 Fang MC, Fan D, Sung SH, *et al*. Validity of using inpatient and outpatient administrative codes to identify acute venous thromboembolism: the CVRN VTE study. *Med Care* 2017;55:e137–43.
26 Delate T, Jones AE, Clark NP, *et al*. Assessment of the coding accuracy of warfarin-related bleeding events. *Thromb Res* 2017;159:86–90.
27 McCormick N, Bhole V, Lacaille D, *et al*. Validity of diagnostic codes for acute stroke in administrative databases: a systematic review. *PLoS One* 2015;10:e0135834.
28 Wright NC, Curtis JR, Arora T, *et al*. The validity of claims-based algorithms to identify serious hypersensitivity reactions and osteonecrosis of the jaw. *PLoS One* 2015;10:e0131601.
29 Liede A, Hernandez RK, Roth M, *et al*. Validation of International Classification of Diseases coding for bone metastases in electronic health records using technology-enabled abstraction. *Clin Epidemiol* 2015;7:441–8.
30 Biggerstaff KS, Frankfort BJ, Orengo-Nania S, *et al*. Validity of code based algorithms to identify primary open angle glaucoma (POAG) in Veterans Affairs (VA) administrative databases. *Ophthalmic Epidemiol* 2018;25:162–8.
31 Thuy Trinh LT, Achat H, Loh SM, *et al*. Validity of routinely collected data in identifying hip fractures at a major tertiary hospital in Australia. *Health Inf Manag* 2018;47:38–45.
32 Strunk A, Midura M, Papagermanos V, *et al*. Validation of a case-finding algorithm for hidradenitis suppurativa using administrative coding from a clinical database. *Dermatology* 2017;233:53–7.
33 Molnar AO, van Walraven C, McArthur E, *et al*. Validation of administrative database codes for acute kidney injury in kidney transplant recipients. *Can J Kidney Health Dis* 2016;3:108.
34 Rowe C, Vittinghoff E, Santos GM, *et al*. Performance measures of diagnostic codes for detecting opioid overdose in the emergency department. *Acad Emerg Med* 2017;24:475–83.
35 Jolley RJ, Quan H, Jetté N, *et al*. Validation and optimisation of an ICD-10-coded case definition for sepsis using administrative health data. *BMJ Open* 2015;5:e009487.
36 Redondo-González O, Tenías JM, Arias Á, *et al*. Validity and reliability of administrative coded data for the identification of hospital-acquired infections: an updated systematic review with meta-analysis and meta-regression analysis. *Health Serv Res* 2018;53:1919–56.