A medição da produção do hospital – a importância da fiabilidade

CARLOS COSTA
PAULO NOGUEIRA

O incremento gradual dos gastos com a saúde tem vindo a suscitar um grande debate sobre o sector.
Ao nível micro, ou seja, dos serviços de saúde, tem vindo a ser apontada a necessidade crescente de se desenvolverem mecanismos efectivos para a avaliação e promoção da eficiência.

1. Introdução

Embora a riqueza das experiências internacionais tenha proporcionado uma grande variedade de alternativas, pode afirmar-se que a sua grande maioria pode ser enquadrada nas seguintes áreas:

- Medição da produção do hospital;
- Garantia da qualidade dos cuidados prestados;
- Avaliação dos custos;
- Financiamento dos hospitais.

Para qualquer destas perspectivas uma questão assume particular importância – a identificação do produto do hospital.

Associadas a esta metodologia – definição do produto do hospital – estão reconhecidas diversas limitações e problemas (Hornbrook, 1982; Costa, 1991). No entanto, no âmbito deste artigo irão ser analisados os aspectos relacionados com a fiabilidade dos instrumentos/medidas que estão ou podem vir a ser utilizados nos hospitais.

Isto porque a reprodutibilidade dos dados e da informação, bem como das medidas e instrumentos que estão a ser utililizados, é uma condição necessária para que todos os agentes económicos associados à actividade hospitalar (proprietários, gestores, prestadores, financiadores e consumidores) tenham confiança na gestão e nos resultados destas organizações.

Em Portugal, embora o debate não tenha ainda atingido proporções nem intensidade idênticas às dos seus parceiros internacionais, existem algumas evidências ao nível operacional que justificam uma discussão sobre a fiabilidade.

Em primeiro lugar, decorrente da crescente atenção exercida sobre a actividade dos hospitais e, concomitantemente, com a avaliação da respectiva efectividade e eficiência.

Em segundo lugar, porque existe um sistema de classificação de doentes aplicado aos hospitais portugueses – os grupos de diagnóstico homogéneos (GDHs).

Finalmente, porque o actual sistema de financiamento dos hospitais portugueses (relativamente ao internamento) – financiamento prospectivo por caso – requer igualmente uma opção prévia sobre a unidade de medida da produção e, consequentemente, a respectiva fiabilidade.

Nesse sentido, no presente artigo irá ser abordada a fiabilidade em sistemas de classificação, essencialmente os aspectos relacionados com: (1) a sua conceptualização; (2) o desempenho de alguns sistemas de classificação de doentes face a este atributo; (3) a operacionalização da sua avaliação.

Em relação aos aspectos referidos, proceder-se-á a uma análise crítica sobre o actual estado da arte e formular-se-ão recomendações para ultrapassar alguns problemas existentes.

2. Fiabilidade: conceptualização e âmbito de aplicação

A fiabilidade pode ser encarada pela probabilidade de uma medida, quando aplicada por diferentes classificadores/utilizadores ou pelos mesmos em diferentes períodos de tempo, produzir os mesmos resultados (Horn e Horn, 1986).

□
Carlos Costa é assistente da disciplina de Economia da Saúde da ENSP; Paulo Nogueira é aluno do mestrado em Probabilidades e Estatística da Faculdade de Ciências e colaborador do Centro de Epidemiologia e Biostatística do Instituto Nacional de Saúde Ricardo Jorge.

Nesse sentido, o âmbito de aplicação da fiabilidade pode assumir três perspectivas (Thomas e Ashcraft, 1989):

- A fiabilidade entre classificadores, ou seja, refere-se à probabilidade de um sistema de classificação ser utilizado da mesma forma por diferentes classificadores;
- A fiabilidade intraclassificador refere-se ao mesmo fenómeno, mas considerando o comportamento do mesmo classificador em diferentes períodos;
- A possibilidade de manipulação, ou seja, a probabilidade de os agentes/classificadores de um determinado hospital classificarem os doentes (por exemplo) em posições mais favoráveis.

Para os dois primeiros aspectos – fiabilidade inter e intraclassificador –, os factores mais decisivos, ou que podem influenciar mais a fiabilidade, são os seguintes (Thomas et al., 1986; Aronow, 1988; Charbonneau et al., 1988; Richards et al., 1988):

- A precisão na definição das categorias e, consequentemente, no grau de ambiguidade em relação às categorias em que cada objecto pode ser classificado;
- O volume de categorias utilizado e, consequentemente, o número de oportunidades para uma má classificação;
- A clareza das instruções em relação ao processo de recolha e de classificação;
- A preparação e uniformização dos classificadores.

Todos estes aspectos (com excepção do último) dizem respeito às características das medidas que estão a ser utilizadas, pelo que implicam a adopção de cuidados especiais no momento da definição operacional das mesmas.

Atendendo a que estas questões irão ser alvo de uma discussão específica no capítulo seguinte – a fiabilidade e os sistemas de classificação de doentes –, remete-se para esta parte do artigo a discussão sobre as principais «armadilhas» associadas com a fiabilidade dos sistemas de medição.

No que se refere à possibilidade de manipulação, os principais problemas decorrem (Thomas et al., 1986):

- Da extensão da inclusão de variáveis/indicadores, na medida em que não pertencem exclusivamente ao fenómeno que se pretende analisar. Por exemplo, se um sistema pretende classificar os doentes do hospital, uma potencial fonte de enviesamento e, concomitantemente, de possibilidade de manipulação decorre da inclusão nessa mesma medida de características que dizem respeito às respectivas opções de tratamento de cada hospital;
- Da objectividade das instruções existentes para a classificação. Ou seja, a comparação entre a existência de instruções ambíguas no respectivo processo de codificação, em que, consequentemente, cada classificador usufrui a oportunidade de poder interpretar cada fenómeno, em oposição a um método de decisão em que as respectivas instruções são precisas e passíveis da mesma interpretação por parte de todos os classificadores;
- Da probabilidade de cada variável/indicador utilizado na medida produzir informação útil e única. A questão em análise neste plano refere-se essencialmente à utilização de indicadores redundantes e de informação supérflua e não pertinente em relação ao fenómeno que se pretende medir.

Mais uma vez se refere que uma discussão mais aprofundada sobre estes aspectos terá lugar no momento de discussão da fiabilidade entre diversos sistemas de classificação de doentes.

3. A fiabilidade e os sistemas de classificação de doentes

Até ao momento inúmeros estudos têm sido realizados comparando a fiabilidade de diversos sistemas de classificação de doentes (Thomas et al., 1986; Thomas e Ashcraft, 1989; Aronow, 1988; Charbonneau et al., 1988; Iezzoni e Moskowitz, 1986; Mullin, 1985).

Embora a metodologia e os resultados não tenham sido idênticos, pode evidenciar-se o seguinte:

- Na generalidade, os sistemas de classificação de doentes que integram o conceito de severidade do estado do doente em função de critérios objectivos – APACHE e MEDISGROUPS (a este propósito consultar Costa, 1991) – apresentam níveis de fiabilidade mais elevados;
- Pelo contrário, os sistemas de medição de severidade que são baseados nos diagnósticos (principal e secundários) – *patient management categories e disease staging* –, bem como os DRGs *(diagnosis related groups)* ou GDHs (grupos de diagnóstico homogéneos), como são conhecidos em Portugal, apresentam problemas na respectiva fiabilidade.

Tem sido apresentada uma grande variedade de argumentos, genéricos – respeitantes à filosofia de cada sistema de classificação – e específicos – relacionados com os princípios e operacionalização de cada sistema de classificação de doentes –, para justificar estas diferenças na respectiva fiabilidade.

Em relação ao primeiro aspecto – enquadramento de cada sistema de classificação de doentes –, os principais problemas/conclusões são os que se seguem.

1. Pode existir uma relação directa entre a fiabilidade de cada sistema de classificação de doentes e os respectivos custos de implementação e de exploração. Ou seja, os sistemas que se baseiam nos *uniform hospital discharge data summary* (UHDDS), ou resumos de alta, e concomitantemente na ICD-9-CM (*International Classification of Diseases – Clinical Modifications*, 9.ª revisão), como os PMC *(disease staging)* ou DRGs, apresentam custos mais baixos do que os sistemas que necessitam de recorrer a elementos constantes dos processos clínicos dos hospitais.

Esta questão – escolha de sistemas de classificação de doentes com menores custos – pode, assim, conduzir a problemas de fiabilidade. Neste sentido, embora os custos de implementação e de exploração de cada sistema devam constituir um critério importante para a respectiva adopção, não deverão constituir um elemento decisivo nesta mesma escolha, e, aspecto mais importante, a análise não deverá ser exercida em termos de «custos mínimos», mas sim na perspectiva «custo-efectividade», sendo, no caso vertente, a efectividade apreciada em função da fiabilidade.

2. A assumpção da validade dos elementos constantes dos processos clínicos não tem sido idêntica para todos os planos de apreciação. De facto, por razões que se discutirão a seguir, existem mais problemas na codificação dos diagnósticos do que na interpretação dos resultados das análises clínicas e dos sinais vitais, pelo que se infere que os sistemas de classificação de doentes que privilegiam a severidade (os baseados em dados objectivos) são, à partida, mais fiáveis.

Na realidade, embora aprioristicamente esta conclusão seja pertinente, será mais aconselhável que se tenham alguns cuidados com a utilização dos resultados das análises clínicas e dos sinais vitais para que desta forma se minimizem os enviesamentos na medição das características dos doentes.

3. A existência de problemas na codificação dos diagnósticos resultantes da utilização da ICD-9-CM. Estes problemas colocam-se a dois níveis: *na escolha do diagnóstico principal e na sobreposição clínica (clinical overlap).*

Em relação à escolha do diagnóstico principal, é facilmente compreensível que dois ou mais classificadores possam escolher diferentes diagnósticos principais.

A este respeito, Doremus e Michenzi (1983), por exemplo, elaboraram um estudo em que os resultados resultantes da recodificação do diagnóstico principal eram diferentes em cerca de 47% dos valores inicialmente apresentados.

Quando tal ocorre, e atendendo à metodologia dos DRGs, por exemplo, poderá igualmente ser atribuído um DRG diferente, e então surgem problemas com a fiabilidade deste sistema de classificação. Os PMC (*patient management categories*) tentam ultrapassar este problema através da imposição de que o diagnóstico principal (aquele que condiciona a escolha do respectivo PMC) será sempre aquele que apresentar um custo mais elevado. No entanto, tal procedimento não é correcto, pois, para além de não se alcançar a fiabilidade do respectivo sistema, introduzem-se ainda problemas na sua validade, tanto de construção como de conteúdo.

Outra questão ainda associada à transposição entre diagnósticos presentes e a atribuição de um determinado escalão de classificação (número do DRG ou do PMC, por exemplo) deriva da quantidade de diagnósticos secundários, visto que é frequente que o volume destes condicione a atribuição da respectiva categoria. Ora, esta questão suscita problemas na fiabilidade decorrentes de diferentes procedimentos entre hospitais e, inclusivamente, através da inclusão de mais um diagnóstico secundário para se atingir uma categoria mais favorável, o que indica claramente uma possibilidade de manipulação dos dados.

A sobreposição clínica na atribuição dos diagnósticos resultantes da aplicação da ICD-9-CM é um problema mais grave, porque é estrutural.

Na realidade, são conhecidos os problemas desta classificação (Thomas e Ashcraft, 1989; Iezzoni e Moskowitz, 1986; Mullin, 1985), designadamente os respeitantes ao facto de a mesma apresentar classes que não são mutuamente exclusivas e classes com um âmbito de aplicação distinto (umas referem-se a problemas de saúde, outras a sintomas, enquanto outras se referem a achados clínicos ou mesmo a indicadores de severidade).

Este problema está igualmente presente nos DRGs, visto que alguns podem ser designados por «grupos relacionados de sintomas» – DRG 143 –, outros por «grupos relacionados de patologia» – DRGs 132 e 133 –, ou mesmo por «grupos relacionados de severidade» – DRG 127 (Iezzoni e Moskowitz, 1986).

Tudo isto conduz à sobreposição clínica – possibilidade de se atribuírem diagnósticos diferentes e, concomitantemente, DRGs diferentes para as mesmas características dos doentes –, em que, portanto, a questão da falta de fiabilidade está presente e que, inclusivamente, implica que este e outros sistemas de classificação não apresentem validade de conteúdo.

4. A existência de diferenças na fiabilidade dentro do mesmo hospital e entre hospitais. De facto, é aceite que potencialmente existem menos problemas na fiabilidade dentro de cada hospital – os procedimentos são mais homogéneos neste plano – do que entre hospitais.

Tal suscita questões e cuidados suplementares, visto que será natural que a possibilidade de ocorrência deste fenómeno (por exemplo, na atribuição do diagnóstico principal e do volume de diagnósticos secundários) implique que se considerem como casos semelhantes (classificados na mesma categoria) doentes que apresentam características diferentes, e vice-versa, o que, para além de enviesar toda a comparação da actividade entre diversos hospitais, constitui igualmente uma limitação na fiabilidade de um determinado sistema de classificação de doentes.

Existem ainda outros problemas associados com a operacionalização de cada sistema de classificação de doentes que são igualmente importantes para a discussão da fiabilidade.

A este propósito existe um excelente artigo de Thomas et al. (1986) em que a fiabilidade dos principais sistemas de classificação de doentes é discutida, pelo que se evita uma duplicação desta análise.

Por outro lado, atendendo a que os DRGs são o sistema de classificação de doentes implementado em Portugal, a exemplificação e a análise crítica irão incidir essencialmente sobre esta medida.

Para delimitar os contornos da discussão irão ser tomados em linha de conta os aspectos enunciados no capítulo anterior – fiabilidade: conceptualização e âmbito de aplicação.

Neste sentido, referem-se os aspectos mais importantes.

5. Precisão na definição das categorias. A metodologia de construção dos DRGs encontra-se bem definida e é actualmente conhecida (Fetter et al., 1980), pelo que se dispensa uma caracterização neste artigo.

Para além dos problemas de validade de construção, de conteúdo e de predição (aspectos que não serão debatidos neste artigo), podem surgir questões que prejudiquem a fiabilidade deste sistema de classificação de doentes.

Embora possam enumerar-se diversas razões para se concluir pela falta de precisão na definição das categorias nos DRGs, o mais importante refere-se à sobreposição clínica na codificação dos diagnósticos.

Neste sentido, a conclusão mais importante a retirar é a seguinte: a precisão na definição das categorias não é somente um atributo do sistema de classificação de doentes em si mesmo, mas deve igualmente ser considerada em termos dos dados e da informação de que este precisa para se desenvolver.

6. No que se refere ao volume de categorias utilizado e à clareza das instruções, os problemas são em tudo semelhantes ao descrito anteriormente, pelo que deve concluir-se que os DRGs apresentam potencialmente problemas de fiabilidade essencialmente derivados do processo de codificação dos diagnósticos.

7. Em relação à preparação e uniformização dos classificadores, deve referir-se que em Portugal, à partida, se tentou controlar esta «variável», visto que foram realizadas diversas acções de formação dos codificadores.

Contudo, atendendo, por um lado, às potenciais diferenças entre fiabilidade dentro de cada instituição e entre os diversos hospitais e, por outro lado, à possibilidade de os hospitais apresentarem comportamentos distintos no processo de escolha do diagnóstico principal, bem como a enumeração dos diagnósticos secundários, inclusivamente para poderem usufruir de ganhos financeiros, seria aconselhável que se realizassem periodicamente estudos de fiabilidade intra e inter-hospitais.

O não cumprimento deste procedimento deve suscitar sérias reservas, podendo, inclusivamente, duvidar-se da validade de utilização deste mecanismo tanto para financiar os hospitais como para comparar a sua actividade.

8. Em relação à pertinência das variáveis incluídas no sistema de classificação, os DRGs apresentam desde logo grandes problemas metodológicos.

Na realidade, a lógica do sistema não considera que para a definição das categorias se utilizem elementos que digam respeito somente às características dos doentes, visto que a opção pelo tratamento («critério de utilização») constitui igualmente um elemento que permite a diferenciação dos produtos.

Este aspecto, para além de aceitar que para um doente ser classificado em determinada categoria não é necessário ter presente a doença, mas somente tenha sido tratado como se ela estivesse presente, permite igualmente que os hospitais escolham o «produto» mais conveniente, estando naturalmente presente a possibilidade de manipulação e, consequentemente, uma menor fiabilidade.

9. O problema da redundância das variáveis incluídas no sistema de classificação de doentes assume uma dupla natureza: em relação à validade do conteúdo e à própria fiabilidade.

Nesta perspectiva, a prática que potencialmente poderá ser utilizada nos DRGs é a inclusão de diagnósticos secundários que permitam a classificação noutra categoria. Nesta eventualidade, mais uma vez se estará na presença de procedimentos que colidem com o princípio da fiabilidade.

Ao longo deste capítulo foram identificados os principais problemas, genéricos e específicos, que podem diminuir a fiabilidade dos sistemas de classificação de doentes, evidenciando-se algumas das limitações mais conhecidas em relação aos DRGs.

Em seguida irão ser debatidos os problemas decorrentes da operacionalização da medição da fiabilidade. Ou seja, qual é a estatística mais adequada para se avaliar a fiabilidade de uma medida?

4. A operacionalização da medição da fiabilidade

A fiabilidade entre classificadores é um conceito estatístico que pode ser avaliado em termos de percentagem de concordância (Horn e Horn, 1986).

Contudo, como a concordância entre os classificadores pode acontecer por acaso, é necessário desenvolver estatísticas que tomem em consideração a concordância por acaso na análise da fiabilidade. Tem sido nesta linha de pensamento que estatísticas como o coeficiente de fiabilidade interclassificadores (*Ri*) e como o *Kappa* de Cohen (K) têm sido desenvolvidas (Thomas e Ascraft, 1989).

Na generalidade, o *Kappa* é uma estatística que indica o grau de concordância entre dois classificadores, mas que ignora o grau de discordância existente, e o *Ri* indica não só o número de discordâncias encontrado, mas o próprio grau de discordância (Thomas et al., 1986). É basicamente por esta razão que estes autores têm defendido a aplicação do coeficiente de fiabilidade interclassificadores (*Ri*) para a avaliação da fiabilidade.

No entanto, o *Ri* trata os valores de cada medida como se os mesmos assumissem os de uma escala de intervalos. Ou seja, é uma escala com significado «quantitativo», em que as distâncias entre dois pontos são idênticas e a que podem aplicar-se as estatísticas paramédicas (Galvão de Melo, 1982).

A realidade é, porém, diferente. Por exemplo, os DRGs constituem uma escala nominal (constituída por um conjunto de classes de equivalência, em que a única medida de tendência central que se pode aplicar é a moda – mesmos autores e obra) e o APACHE ou os MEDISGROUPS constituem uma escala ordinal, em que os valores são classificados em classes com ordenações, mas em que as diferenças entre dois pontos não são semelhantes e em que a estatística que melhor representa a distribuição é a mediana (mesma obra).

Daqui pode concluir-se que, atendendo às características dos sistemas de classificação de doentes em aplicação nos hospitais, a utilização do *Ri* não é a mais adequada, visto que os respectivos resultados podem vir enviesados.

Uma boa ilustração deste princípio é encontrada no estudo de Thomas e Ascraft (1989), em que, para além de aplicar uma estatística paramétrica a dados não paramétricos, foi necessário recodificar as escalas originais de cada sistema de classificação de doentes.

Para os DRGs o procedimento seguido pelos autores foi o seguinte: utilizou-se o peso específico de cada DRG em detrimento do respectivo número, introduzindo-se posteriormente escalões de variação.

Nesta situação facilmente se compreende a falta de racionalidade da metodologia, pois os codificadores podiam ter atribuído um diagnóstico principal e um número de diagnósticos secundários diferentes, que conduziriam a DRGs diferentes, mas, desde que aos mesmos estivesse associado um peso específico semelhante, considerava-se que estavam de acordo, o que, como se observa, não é verdade.

Assim, deve concluir-se pela não viabilidade de utilização do coeficiente de fiabilidade interclassificadores (*Ri*), pelo que de seguida irão debater-se os aspectos metodológicos relacionados com o *Kappa* de Cohen (*K*).

A estatística *Kappa* foi originalmente proposta numa base *ad hoc* como uma estatística descritiva indicadora do grau de concordância, para além do acaso, entre duas classificações por objecto ou sujeito (resposta dicotómica) (Bloch e Kraemer, 1989).

Assim, o coeficiente *Kappa* era uma estatística associada a uma matrix 2 × 2.

Observe-se a *Tabela I*.

Sendo n_{ij} o número de «objectos» colocados na categoria i por B e na categoria j por A, $i, j = 1, 2$, $n_{i\bullet} = n_{i1} + n_{i2}$, $n_{\bullet j} = n_{1j} + n_{2j}$, $n = \sum_{ij} \Sigma n_{ij}$.

Definindo-se a porporção observada de acordos por

$$p_0 = \frac{\sum_i n_{ii}}{n_{..}} = \frac{n_{11} + n_{22}}{n_{..}}$$

aprioristicamente, esta definição não apresenta obstáculos de maior, visto que os classificadores só estão de acordo quando classificam o mesmo objecto na mesma classe. Mas parte dessa concordância observada deve-se ao acaso, visto que existe a hipótese de os classificadores concordarem sem que tal se deva única e exclusivamente à característica do objecto ou do sujeito.

Tabela I
Formato para o cálculo do *Kappa* de Cohen

		Classificador A		
		1	2	
Classificador *B*	1	n_{11}	n_{12}	$n_{1\bullet}$
	2	n_{21}	n_{22}	$n_{2\bullet}$
		$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{..}$

Para se ultrapassar esta questão define-se a proporção esperada de acordos devidos ao acaso por

$$p_e = \frac{\sum_i n_{i.} n_{.i}}{n_{..}^2} = \frac{n_{1.}n_{.1} + n_{2.}n_{.2}}{n_{..}^2}$$

Pode então corrigir-se a porporção de concordância observada pela proporção de concordância esperada, isto é, $p_0 - p_e$.

Esta medida assume o valor zero quando os classificadores concordam apenas por acaso. No entanto, quando ocorre perfeita concordância ou discordância, a medida $p_0 - p_e$ não assume os valores + 1 e 1, respectivamente, como seria desejável.

Para que tal aconteça é necessário fazer uma transformação, em que se define

$$\hat{K} = \frac{p_0 - p_e}{1 - p_e}$$

obtendo-se, assim, a estatística *Kappa*, como uma medida de concordância corrigida para o acaso.

Após a discussão do K para uma matriz 2×2, interessa proceder à sua generalização para m categorias. Este procedimento não é complexo, devendo proceder-se como se exemplifica na *Tabela II*.

Analogamente, pode definir-se

$$\hat{K} = \frac{p_0 - p_e}{1 - p_e}$$

Tabela II
Formato para o cálculo do *Kappa*

		Classificador B					
		1	2	3	...	m	
Classificador A	1	n_{11}	n_{12}	n_{13}	...	n_{1m}	$n_{1.}$
	2	n_{21}	n_{22}	n_{23}	...	n_{2m}	$n_{2.}$
	3	n_{31}	n_{32}	n_{33}	...	n_{3m}	$n_{3.}$
	‡	‡	‡	‡	‡	‡	‡
	m	n_{m1}	n_{m2}	n_{m3}	...	n_{mm}	$n_{m.}$
		$n_{.1}$	$n_{.2}$	$n_{.3}$	...	$n_{.m}$	$n_{..}$

Tabela III

Estatística *Kappa*	*Nível de concordância*
< 0.00	Pobre
0.00 – 0.20	Ligeiro
0.21 – 0.40	Razoável
0.41 – 0.60	Moderado
0.61 – 0.80	Substancial
0.81 – 1.00	Quase perfeito

Com p_0 – proporção de concordância observada:

$$p_0 = \frac{\sum_i n_{ii}}{n_{..}} = \frac{n_{11} + n_{22} + \ldots + n_{mm}}{n_{..}}$$

e p_e – proporção de concordância esperada por acaso:

$$p_e = \frac{\sum_i n_{i.} n_{.i}}{n_{..}^2} = \frac{n_{1.}n_{.1} + n_{2.}n_{.2} + \ldots + n_{m.}n_{.m}}{n_{..}^2}$$

as proporções (pij) podem ser obtidas após a divisão por $n_{..}$ dos elementos constantes da *Tabela II*.

Donde surge:

$$p_0 = \sum_i p_{ii} \text{ e } p_e = \sum_i p_{i.} p_{.i}$$

Fleiss (1981) mostra que a generalização do *Kappa* pode reduzir-se a m estudos de matrizes 2×2, em que a cada passo se isola a classe em estudo e se colapsam as restantes numa só classe, surgindo o valor de *Kappa* da matriz $m \times m$ como a soma das diferenças individuais $P_0 - P_e$ (isto é, os numeradores dos *kappas* individuais) a dividir pela soma das diferenças individuais $1 - P_e$ (isto é, os denominadores dos *kappas* individuais).

Neste sentido, o *Kappa* total é obtido através da média ponderada dos *kappas* individuais para cada tabela individual, isto é:

$$\hat{K} = \frac{\sum_i k_i (1 - P_e^{(i)})}{\sum_i (1 - P_e^{(i)})}$$

No entanto, subsistem ainda alguns problemas com a aplicação da estatística *Kappa*, nomeadamente os relacionados com a dificuldade de interpretação da sua magnitude, com a relação entre *Kappa* e prevalência e com a discussão entre concordância e associação.

Em relação ao primeiro aspecto – dificuldade de interpretação –, Landis e Koch (1977) construíram uma tabela (*Tabela III*) que, actualmente, é considerada como *standard* para se apreciar o nível de concordância.

Contudo, como a amplitude *Kappa* depende do número de objectos a classificar e por razões associadas ao desconhecimento da distribuição de *Kappa*, esta tabela não deve ser aceite (Barlow e Azem, 1991; Nogueira, 1993).

No que respeita ao segundo aspecto, sabe-se que o valor de *Kappa* depende da prevalência da característica em estudo quando os parâmetros de sensibilidade e especificidade são fixos, sendo a sensibilidade a proporção dos objectos que verdadeiramente possuem a característica e que como tal são classificados e especificidade a proporção dos objectos que realmente não possuem a característica e que como tal são classificados.

Finalmente, deve referir-se que, na realidade, *Kappa* não é uma medida de concordância, mas somente de associação (Bloch e Kraemer, 1989), visto que a consideração dos valores do acaso afecta esta característica (Brennan e Prediger, 1981).

O *Kappa* de Cohen apresenta a seguinte forma genérica:

$$k(r) = w(r)\,k(1) + [1 - w(r)]\,k(0)$$

onde

$$w(r) = \frac{P(1-Q)\,r}{P(1-Q)\,r + (1-P)\,Q(1-r)}$$

com:

- r = índice que reflecte a importância relativa das classificações na diagonal principal (*Tabela I*):

$$r = \frac{n_{11} - n_{12}}{n_{11} - n_{12} - n_{21} - n_{22}}$$

- P, $(1-P)$ distribuição marginal do classificador B;
- Q, $(1-Q)$ distribuição marginal do classificador A.

Estas distribuições marginais surgem tal como acontece num contexto de associação, e não num de concordância, em que estas distribuições são supostamente iguais (Bloch e Kraemer, 1989; Nogueira, 1993).

Para se ultrapassar este problema, ou seja, para se obter uma medida de concordância, Bloch e Kraemer (1989) propõem o desenvolvimento de uma nova estatística – o *Kappa* intraclasses –, em que apenas a estrutura do acaso é diferente do *Kappa* de Cohen.

Noutro estudo (Nogueira, 1993) demonstra-se que o *Kappa* intraclasses dá origem a uma estrutura de discordância mais clara do que o coeficiente de Cohen, sendo os seus valores sempre inferiores ou iguais aos de K de Cohen, pelo que esta estatística é claramente mais conservativa, tal como é sugerido por Guggenmoos-Holzmann (1993).

No entanto, embora o *Kappa* intraclasses seja uma medida de concordância melhor do que o *Kappa* de Cohen, as dificuldades para o seu cálculo, bem como a sua semelhança assintótica com K, não abrem ainda grandes perspectivas para a solução do problema da interpretação dos valores obtidos e da real quantificação dos valores de concordância.

Neste sentido avultam ainda os problemas de qualificar e quantificar a concordância, pelo que estes instrumentos (K de Cohen e *Kappa* intraclasses) devem ser considerados necessários, mas não suficientes, para se avaliar a fiabilidade das medidas.

É com este propósito que, de seguida, se apresentarão os modelos loglineares, argumentando-se que os mesmos constituem uma forma expedita e efectiva para complementar a apreciação facultada pela estatística *Kappa* e, concomitantemente, para se avaliar a fiabilidade.

Os modelos loglineares partem do pressuposto de que os classificadores raramente conseguem atingir concordância perfeita, essencialmente porque as suas percepções sobre o significado das categorias são diferentes.

Para a sua aplicação deve adicionar-se ao modelo tradicional de independência uma componente que descreve a associação intrínseca às classificações por categorias (*baseline association between classifications*) e uma componente relativa à diagonal principal que representa a incidência adicional de perfeita concordância (para esclarecimentos adicionais, cf. Nogueira, 1993).

A conjugação de todos estes factores permite a obtenção de vários modelos, entre os quais:

a) Modelo de independência:

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B$$

onde:

- m_{ij} denota o número de «sujeitos» ou «objectos» classificados na célula (i, j);
- μ denota um nível à volta do qual os classificadores classificam;
- λ_i^A reflecte a probabilidade de o classificador A classificar na classe i;
- λ_j^B reflecte a probabilidade de o classificador B classificar na classe j.

Assim, o modelo resulta do facto de, sob a hipótese de independência, a probabilidade conjunta ser o produto das probabilidades de cada classificador, logo a soma dos respectivos logaritmos;

b) Modelo de extraconcordância uniforme:

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \delta(i,j) = \begin{cases} \delta, & i = j \\ 0 & \text{C C} \end{cases}$$

com

$$\delta_{(i,j)}$$

Ao modelo descrito em *a*), de independência, junta-se uma componente de associação, $\delta(i, j)$, relativa à incidência na diagonal principal, já que é esta incidência que representa a extraconcordância. Esta componente é uniforme para toda a diagonal principal.

A designação de «diagonal principal» é utilizada com o sentido matemático, significando que os classificadores classificam o mesmo sujeito ou objecto numa mesma classe, ou seja, quando $i = j$;

c) Modelo de quase independência ou de extraconcordância não uniforme:

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \delta(i, j)$$

com

$$\delta(i, j) = \begin{cases} \delta, i = j, i = 1, \ldots, r \\ 0 \quad C\,C \end{cases}$$

em que r representa o número de classes em jogo. Como implicitamente assumido nestes modelos, este número é sempre maior do que 2.

Este modelo contém igualmente uma componente relativa à diagonal principal, diferindo do anterior no facto de permitir níveis diferentes para cada elemento da diagonal principal, apresentando, assim, o inconveniente de ser saturado na diagonal principal:

d) Associação uniforme:

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \beta u_i u_j$$

em que u_i são scores fixos para as r classes.

Acrescenta-se, assim, ao modelo de independência uma componente de associação, designada linear × linear, que permite a compreensão da associação devida às diferentes interpretações e percepções dos classificadores em relação às classes sujeitas à classificação;

e) Extraconcordância com associação quase uniforme:

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \beta_{ij} + \delta(i, j)$$

com

$$\delta(i, j) = \begin{cases} \delta, i = j \\ 0 \quad C\,C \end{cases} \quad e \quad \beta_{ij} = \beta u_i u_j$$

A combinação dos dois modelos apresenta como principal vantagem o facto de disponibilizar uma maior compreensão da concordância observada, mantendo como principal desvantagem a saturação na diagonal principal;

f) Extraconcordância com associação uniforme:

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \beta_{ij} + \delta(i, j)$$

com

$$\delta(i, j) = \begin{cases} \delta, i = j \\ 0 \quad C\,C \end{cases} \quad e \quad \beta_{ij} = \beta u_i u_j$$

Tal como no modelo *e*), este resulta de uma combinação de dois modelos, dando-se aqui uma maior liberdade à diagonal principal.

O modelo que teoricamente se considera melhor para se apreciar a fiabilidade é o de extraconcordância.

Isto porque apresenta os seguintes aspectos positivos:

- Utiliza uma ordem das categorias;
- Não assume que as classificações são independentes quando os classificadores discordam;
- Não é saturado na diagonal principal;
- Os parâmetros de associação intrínseca e de extraconcordância na diagonal principal são de fácil interpretação;
- Trata-se de um modelo de quase-simetria;
- É um modelo que explica razoavelmente bem a parte da variação suplementar em relação ao modelo de independência;
- Apresenta ainda uma relação estrita com os *odd-ratios* através daquilo que se designa por *category distinguis hability* (para aprofundamento, consultar Agresti, 1988, e Becker, 1989).

A metodologia de decisão é relativamente simples, embora faseada. Após o cálculo do valor de *Kappa*, e na eventualidade de obtenção de um valor razoável, deve proceder-se ao cálculo dos diferentes modelos loglineares, podendo utilizar-se um programa informatizado de estatística (o GLIM, por exemplo).

Quando, por exemplo, os resultados práticos de um determinado estudo demonstram que o modelo que melhor se ajusta (fornece uma melhor explicação dos dados) é o de associação uniforme, e estando-se na presença de valores significativos de *Kappa*, então deve concluir-se que a concordância observada é devida a uma razoável compreensão por parte dos classificadores em relação à escala utilizada.

Mas, quando não se está na presença de extraconcordância, os responsáveis pela realização e avaliação do estudo devem questionar-se sobre a própria medida.

Ou seja, resultados desta natureza (sempre que o modelo que melhor ajuste os dados seja outro que não o de extraconcordância) implicam a necessidade de se melhorar a escala de medição (imprimir maior objectividade) ou de se desenvolverem novas acções de formação tendentes a homogeneizar os procedimentos entre classificadores (diminuir diferenças resultantes do comportamento dos classificadores).

Em síntese, deve afirmar-se que a avaliação da fiabilidade tem apresentado alguns problemas de natureza estatística.

A utilização de estatísticas, como o coeficiente de fiabilidade interclassificadores (*Ri*), manifesta-se inadequada para o efeito pretendido, essencialmente porque os sistemas de classificação de doentes desenvolvidos até ao momento não se referem a escalas de intervalos.

Por outro lado, a informação disponibilizada pelo *Kappa*, embora seja útil, é insuficiente: este coefi-

ciente é mais uma estatística de associação do que de concordância e é de interpretação difícil.

Assim, deve desenvolver-se uma metodologia faseada, em que após o cálculo do *Kappa* devem ser feitos os apuramentos respeitantes aos modelos loglineares, sabendo-se de antemão que, quando entre as diversas alternativas os resultados do modelo de extraconcordância não forem os que melhor se ajustam, subsistem ainda algumas dúvidas sobre a fiabilidade da medida em estudo.

5. Conclusões

Neste artigo foi evidenciada a necessidade de se desenvolverem mecanismos efectivos para a avaliação e para a promoção da eficiência dos hospitais.

Referiu-se igualmente que, pese embora o facto de este objectivo percorrer áreas tão distintas como a garantia de qualidade, a avaliação dos custos e o financiamento, o desenvolvimento de um sistema de medição da produção constitui, de facto, um elemento fundamental para a apreciação da actividade dos hospitais.

Nesta perspectiva foi amplamente discutida a necessidade de se implementarem metodologias que propiciem a avaliação da fiabilidade dos sistemas de medição da produção do hospital.

Isto porque, para além de constituir uma via que propicia o aumento da confiança por parte de todos os agentes económicos associados com a actividade hospitalar em relação ao seu desempenho, disponibiliza igualmente instrumentos que conferem um maior rigor na prossecução desta finalidade.

Em seguida, referiram-se os aspectos relacionados com a conceptualização e o âmbito de aplicação do conceito «fiabilidade» à medição da produção do hospital, evidenciando-se a fiabilidade inter e intraclassificadores, bem como a possibilidade de manipulação da informação.

A discussão deste conceito – fiabilidade –, em função dos sistemas de classificação de doentes ou de identificação de produtos, foi essencialmente centrada nos *diagnosis related groups* (DRGs), atendendo ao facto de os mesmos se encontrarem aplicados nos hospitais portugueses.

Assim, foram identificados os aspectos mais limitativos, designadamente os relacionados com a «sobreposição clínica», com a pertinência e redundância das variáveis incluídas neste sistema e ainda com a possibilidade de manipulação.

Posteriormente foi discutido o problema da avaliação estatística da fiabilidade.

Após a identificação, caracterização e análise crítica das principais estatísticas a utilizar, foi sugerida uma estratégia de actuação para a sua concretização.

Genericamente, preconizou-se uma actuação faseada, em que os resultados do coeficiente K de Cohen devem ser conjugados com os dos modelos loglineares para, assim, se avaliar e conhecer a amplitude de variação da fiabilidade.

Finalmente, faz-se um apelo para a realização de estudos de investigação sobre a fiabilidade das medidas utilizadas no sector da saúde, essencialmente no que respeita aos DRGs, os quais devem ser desenvolvidos, quer a nível local, quer a nível central, para que dessa forma se garanta uma maior efectividade na avaliação da actividade dos hospitais portugueses.

□ Agradecimentos

Os autores querem exprimir o seu agradecimento ao Prof. Galvão de Melo pelas sugestões e comentários que realizou, especialmente no que se refere ao capítulo «A operacionalização da medição da fiabilidade».
Quaisquer erros ou omissões são, todavia, da inteira responsabilidade dos autores.

□ Bibliografia

AGRESTI, A.
A model for agreement between ratings on an ordinal scale.
«Biometrics», 44, 1988, 539-548.

ARONOW, D. B.
Severity of illness measurement: applications in quality assurance and utilization review.
«Medical Care Review», 45, 1988, pp. 339-366.

BARLOW, W., et. al.
Comparison of methods for calculating a stratified Kappa.
«Statistics in Medicine», 10, 1991, 1465-1472.

BECKER, M. P.
Using association models to analyse agreement data: two examples.
«Statistics in Medicine», 8, 1989, 1199-1207.

BLOCH, D. A., e KRAEMER, H. C.
2 × 2 Kappa coefficients: measures of agreement or association.
«Biometrics», 45, 1989, pp. 269-287.

BRENNAN, R. L., e PREDIGER, D. J.
Coefficient of Kappa: some issues, misuses and alternatives.
«Educational and Psychological Measurement», 41, 1981.

CHARBONNEAU, C., et. al.
Validity and reliability in alternative patient classification systems.
«Medical Care», 26 (8), 1988, pp. 800-813.

COSTA, C.
A severidade da doença – identificação e caracterização de alguns sistemas de classificação.
«Revista Portuguesa de Saúde Pública», 9 (1), 1991, pp. 37-44.

DOREMUS, H. D., e MICHENZI, E. M.
Data quality: an illustration of its potential impact upon a diag-

nostic related group's case mix index and reimbursment.
«Medical Care», 21, 1983, pp. 1001-1010.

FETTER, R. B., et. al.
Case mix definition by diagnostic related groups.
«Medical Care», 19 (2), 1980, suppl.

FLEISS, J. L.
Statistical methods for rates and proportions. Wiley & Sons, 1981.

GALVÃO DE MELO, F.
Modelos e métodos estatísticos em psicologia. Lisboa: ENSP, «Obras avulsas», 1982.

GUGGENMOOS-HOLZMANN, I.
How reliable are chance-corrected measures of agreement?
«Statistics in Medicine», 12 (23), 1993, pp. 2191-2205.

HORN, S. D., e HORN, R. A.
Reliability and validity of the severity of illness index.
«Medical Care», 24, 1986, pp. 159-178.

HORNBROOK, M. C.
Hospital case mix: its definition, measurement and use, part I, The conceptual framework.
«Medical Care Review», 39, 1982, pp. 1-43.

HORNBROOK, M. C.
Hospital case mix: its definition, measurement and use, part II, Review of alternative measures.
«Medical Care Review», 39, 1982, pp. 73-123.

IEZZONI, L. I., e MOSKOWITZ, M. D.
Clinical overlap among medical diagnosis related groups.
«JAMA», 255 (7), 1986, pp. 927-929.

KRAEMER, H. C.
Extension of the Kappa coefficient.
«Biometrics», 36, 1979, 207-216.

LANDIS, J. R., e KOCH, G. G.
The measurement of observed agreement for categorical data.
«Biometrics», 33, 1977, pp. 139-174.

MULLIN, R. L.
Diagnosis related groups and severity – ICD-9-CM, the real problem.
«JAMA», 254 (9), 1985, pp. 1208-1210.

NOGUEIRA, P.
Um estudo sobre concordância para avaliação da validade e fiabilidade de um sistema de classificações nominais. ENSP e Faculdade de Ciências da Universidade de Lisboa, 1993.

RICHARDS, T., et. al.
Measuring differences between teaching and nonteaching hospitals.
«Medical Care», 26 (5), 1988, suppl., pp. s1-s141.

THOMAS, J. W., et. al.
An evaluation of alternative severity of illness measures for use by university hospitals. Ann Arbor, DHSMP, University of Michigan, 1986.

THOMAS, J. W., e ASHCRAFT, M. L. F.
Measuring severity of illness: a comparison of interrate reliability among severity methodologies.
«Inquiry», 26, 1989, pp. 483-492.

□ Résumé

ÉVALUATION DE LA PRODUCTION DE L'HÔPITAL – L'IMPORTANCE DE LA CONFIANCE

L'évaluation de la production de l'hôpital est très importante pour l'acquisition des instruments qui possibiliteront une plus grande eficacité de la qualité des services, de l'évaluation des coûts de l'établissement de systèmes de financement.

Un des aspects cruciels pour l'appréciation des systèmes d'évaluation de la production des hôpitaux est la confiance qu'ils méritent.

En réalité, la probabilité q'une mesure, quand elle est apliquée par différents classificateurs ou par les mêmes mais dans des périodes de temps différentes, puisse produire les mêmes résultats (confiance) est une condition indispensable pour que la confiance dans la gestion et dans l'évaluation des résultats des hôpitaux soit plus grande.

Ainsi, cet article débat les aspects relationés avec la conception et le champ d'application de la confiance, bien que son association avec les systèmes de classification de malades.

Les statisques qui sont plus utilisées à niveau international seront identifiées et charactérisées dans une autre occasion, étant données les dificultés de leur évaluation.

Les raisons principales pour ne pas utiliser le coefficent de confiance interclassificateurs, qui est une statistique utile pour les mesures qui présentent des échelles d'intervalles, sont presentées.

Ensuite, aprés une discussion du Kappa de Cohen, une stratégie est sugérée où les résultats de ce coefficient doivent être combinés avec les modèles loglineaires, afin de connaître l'amplitude de la concordance et aussi sa signification.

Finalement, la nécessité de développer des études sur cette matière, appliqués aux hôpitaux portugais, pour l'évaluation de la situation existante et de l'introduction des altérations convenables, est mise en évidence.

□ Summary

EVALUATING HOSPITAL PRODUCTION – THE IMPORTANCE OF RELIABILITY

To evaluate hospital production is very important for the acquisition of instruments which may enable a bigger effectiveness to guarantee the quality of the services, evaluate the costs and establish financing systems.

One of the crucial aspects for the evaluation of hospital production is its reliability.

The probability that a given measure, when applied by different classifiers or by the same classifier but in a different period of time, produces the same results (reliability) is, in fact, an essential condition to strenghten the confidence in the management and the evaluation of hospital results.

This article will, thus, debate the aspects related to the conceptualization and field of application of reliability, as well as its association with patient classification systems.

Due to the operational difficulties of its evaluation, the statistics internationally more used will be identified and characterized in a different document.

The main arguments for not using the interclassifier reliability coefficent, a useful statistics for measurements with intermission scales, will be presented.

After discussing Cohen's Kappa, a strategy is suggested in which the results of this coefficient should be combined with those of the loglinear models, in order to know not only the extent of the agreement, but also its significance.

The need to develop studies in this area, applied to Portuguese hospitals, as a mean to evaluate the existing situation and to introduce the convenient changes, becomes, eventually, evident.