

Fidedignidade de indicadores de qualidade do cuidado de enfermagem: testando a concordância e confiabilidade interavaliadores¹

Dagmar Willamowius Vituri²
Yolanda Dora Martinez Évora³

Objetivo: testar a concordância e confiabilidade interavaliadores de quinze indicadores de qualidade da assistência de enfermagem. **Método:** pesquisa quantitativa, metodológica, experimental e aplicada, realizada em um hospital terciário, de grande porte, de ensino público do Estado do Paraná. No tratamento dos dados foi utilizada a estatística Kappa para análise das variáveis categóricas – indicadores de 1 a 11 e 15; e o coeficiente de correlação interclasses para as variáveis contínuas – indicadores 12, 13 e 14, considerando-se o intervalo de confiança de 95%. Os dados categóricos foram tratados utilizando-se o software Lee, do Laboratório de Epidemiologia e Estatística do Instituto Dante Pazzanese de Cardiologia – Brasil e os dados contínuos pelo software BioEstat 5.0. **Resultados:** os resultados pela aplicação da estatística Kappa demonstraram excelente grau de concordância, estatisticamente significativa; os valores obtidos para o coeficiente de correlação interclasses denotaram reprodutibilidade/concordância excelente e estatisticamente significativa para os indicadores avaliados. **Conclusão:** os indicadores apresentam excelente confiabilidade e reprodutibilidade, o que evidencia que o desenvolvimento de instrumentos de avaliação válidos e fidedignos é possível, além de imprescindível para gerenciamento da assistência de enfermagem.

Descritores: Enfermagem; Auditoria de Enfermagem; Indicadores de Qualidade em Assistência à Saúde; Estudos de Validação como Assunto; Reprodutibilidade dos Testes.

¹ Artigo extraído da tese de doutorado "Avaliação como princípio da Gestão da Qualidade Total: testando a confiabilidade interavaliadores de indicadores de qualidade da assistência de enfermagem" apresentada à Escola de Enfermagem de Ribeirão Preto, Universidade de São Paulo, Centro Colaborador da OMS para o Desenvolvimento da Pesquisa em Enfermagem, Ribeirão Preto, SP, Brasil.

² Doutoranda, Escola de Enfermagem de Ribeirão Preto, Universidade de São Paulo, Centro Colaborador da OMS para o Desenvolvimento da Pesquisa em Enfermagem, Ribeirão Preto, SP, Brasil. Enfermeira, Hospital Universitário, Universidade Estadual de Londrina, Londrina, PR, Brasil.

³ PhD, Professor Titular, Escola de Enfermagem de Ribeirão Preto, Universidade de São Paulo, Centro Colaborador da OMS para o Desenvolvimento da Pesquisa em Enfermagem, Ribeirão Preto, SP, Brasil.

Endereço para correspondência:

Dagmar Willamowius Vituri
Universidade Estadual de Londrina. Hospital Universitário
Gerência de Riscos
Av. Robert Koch, 60
Vila Operária
CEP: 86038-350, Londrina, PR, Brasil
E-mail: dagmar@uel.br

Copyright © 2014 Revista Latino-Americana de Enfermagem

Este é um artigo de acesso aberto distribuído sob os termos da Licença Creative Commons Atribuição-Não Comercial (CC BY-NC).

Esta licença permite que outros distribuam, editem, adaptem e criem obras não comerciais e, apesar de suas obras novas deverem créditos a você e ser não comerciais, não precisam ser licenciadas nos mesmos termos.

Introdução

A preocupação com a qualidade dos produtos e serviços tem sido objeto de apreensão nos diversos tipos de organização, principalmente no setor saúde, pois, além da influência direta desses serviços sobre a economia, seus clientes estão cada vez mais exigentes quanto ao atendimento de suas necessidades com qualidade⁽¹⁾.

Um grande aliado das organizações de saúde na busca pela qualidade das ações é o enfermeiro e, nesse sentido, existe expectativa dos gestores quanto ao papel desse profissional no gerenciamento do cuidado nas instituições hospitalares, pois, uma das características da prática do cuidar em enfermagem é o contato direto com o cliente, que possibilita conhecer suas necessidades e expectativas⁽¹⁾, além do fato de a profissão ter papel fundamental no resultado da atenção.

Gerenciar implica conhecer o desempenho da instituição em função de seu objetivo e em relação às suas metas⁽²⁾ e isso requer a implementação de sistemas de avaliação e indicadores adequados, que permitam a (re)formulação de diretrizes⁽³⁾. Contudo, a quantificação da qualidade da atenção em saúde tem sido um grande desafio, que se inicia com a decisão do que medir e, depois, encontrar medidas estatísticas de qualidade⁽⁴⁾.

Indicador é uma unidade de medida de uma atividade, que pode ser utilizado para mensurar qualidade e quantidade em organizações de saúde, avaliando aspectos de estrutura, processos e resultados⁽⁵⁻⁶⁾. A possibilidade de desenvolvê-los é ilimitada; porém, a dificuldade está em encontrar um indicador de alta validade para o domínio em estudo⁽⁷⁾, ou seja, para o aspecto do cuidado a ser medido.

O uso de medidas válidas e fidedignas propicia monitorar a qualidade do cuidado dispensado aos pacientes, identificar situações de riscos evitáveis, bem como subsidiar o planejamento de ações corretivas, além de direcionar estratégias e reajuste de metas por meio de ações educativas e de valorização profissional. Ressalta-se que o uso de instrumentos válidos e fidedignos pode contribuir para o avanço do conhecimento da profissão e para o desenvolvimento da teoria que a sustenta⁽⁸⁾.

Fidedignidade implica confiabilidade e concordância, que são aspectos importantes no desenvolvimento de instrumentos de medida. Seus resultados fornecem informações sobre a quantidade de erro inerente a essa, o que reflete a qualidade da medição⁽⁹⁾.

Concordância é o grau em que os escores ou classificações são idênticos⁽⁹⁾ e confiabilidade o grau em que o resultado medido reflete o resultado verdadeiro, ou seja, o quanto uma medida está livre da variância dos erros aleatórios⁽¹⁰⁾. Pode ainda ser definida como a

proporção da variância nos escores de medição, que se devem às diferenças no escore verdade, não ao erro aleatório⁽¹¹⁻¹²⁾. Sendo assim, a confiabilidade estima a consistência e estabilidade da medida e aumenta à medida que o componente de erro diminui⁽¹¹⁾.

Considerando-se a preocupação histórica da enfermagem com a busca pela qualidade da atenção, assinalada desde as ações de cuidado de Florence Nightingale, bem como pela crise de credibilidade atualmente associada aos serviços de saúde brasileiros e pelo potencial do profissional enfermeiro em contribuir para a mudança dessa realidade, por meio da mensuração da qualidade do cuidado prestado, fica evidente a necessidade de que as medidas da qualidade da assistência sejam confiáveis.

O desenvolvimento de instrumentos avaliativos é complexo e objeto de estudo de várias disciplinas. No contexto da prática da enfermagem, observam-se muitas críticas quanto ao seu uso em função de se pretender mensurar construtos e conceitos considerados abstratos e subjetivos.

Acredita-se, contudo, que essas críticas se devam à falta de conhecimento sobre o processo envolvido na concepção e validação de medidas avaliativas, o que resulta em desmotivação para o seu desenvolvimento e, conseqüentemente, impede o avanço da ciência na área.

Diante desse cenário e dando continuidade ao estudo que se iniciou com a validação de conteúdo de indicadores de qualidade da assistência de enfermagem⁽¹³⁾, tem-se, como questão de pesquisa: o instrumento constituído por indicadores de qualidade de cuidados básicos de enfermagem atende ao requisito de confiabilidade para avaliação da qualidade da assistência dispensada a pacientes adultos, internados em unidade médico-cirúrgica de um hospital público de ensino do norte do Estado do Paraná?

A presente pesquisa tem como objetivo testar a concordância e confiabilidade interavaliadores de quinze indicadores de qualidade da assistência de enfermagem.

Metodologia

Trata-se de pesquisa quantitativa, metodológica, experimental e aplicada, realizada em três fases (validação de face, teste-piloto e teste de confiabilidade), desenvolvida em um hospital terciário de grande porte, de ensino público do Estado do Paraná, Brasil. Ressalta-se que, no presente estudo, o foco da apresentação dos resultados e discussão é o teste de confiabilidade da medida proposta.

O estudo iniciou-se com a validação de face de um instrumento composto por quinze indicadores de qualidade, considerando o percentual mínimo de concordância⁽¹⁴⁾ de 80%, com base no instrumento proposto por Vituri⁽¹³⁾, validado pela estratégia de validação de conteúdo por *experts* e reformulado com base nas necessidades sentidas por meio de sua aplicação rotineira na prática diária.

Para o procedimento de validação de face selecionou-se uma amostra intencional de juízes, estagiários do serviço de avaliação da qualidade da assistência de enfermagem do hospital em estudo. Esse serviço funciona como campo de estágio curricular não obrigatório para alunos do terceiro e quarto anos do curso de graduação em enfermagem, que desenvolvem atividades de auditoria operacional e retrospectiva, avaliando a qualidade da assistência de enfermagem, por meio da aplicação sistemática do instrumento de Vituri⁽¹³⁾.

O teste-piloto do instrumento de medida foi realizado por uma amostra intencional de três juízes, sendo dois enfermeiros da instituição (um assistencial e outro gerente em nível de direção) mais a pesquisadora. Esse número foi definido por meio dos estudos de Crocker, Llabre e Miller⁽¹⁵⁾, pois, quanto maior o número de juízes maior a heterogeneidade e, conseqüentemente, menor a confiabilidade e concordância⁽¹⁶⁾.

A compreensão dos indicadores, metodologia de avaliação e aplicabilidade do instrumento foi testada por meio da avaliação de uma amostra aleatória de quinze pacientes adultos, internados em unidade de clínica médica e cirúrgica.

Para determinação da fidedignidade, dentre as três categorias de estimação de concordância (estimativa de consenso, estimativa de consistência e estimativa de avaliação/medição), utilizou-se a estimativa de consenso, que tem como base a suposição de que deva existir perfeita concordância entre os juízes no uso de escalas de pontuação para comportamentos observáveis, o que caracteriza a mesma compreensão do construto⁽¹⁷⁾.

O método utilizado foi o de equivalência, que é usado para avaliação da confiabilidade mediante a aplicação do mesmo instrumento por observadores diferentes, para medir os mesmos fenômenos (confiabilidade interavaliadores ou precisão do avaliador)⁽¹¹⁾.

Esse tipo de teste é indicado, principalmente, para instrumentos clínicos, em que se depende do julgamento do avaliador. Sendo assim, é fundamental estimar a precisão pela correlação dos resultados obtidos de cada sujeito, pois é evidente a variância existente entre os avaliadores⁽¹¹⁾.

A concordância e a fidedignidade dos indicadores foram testadas pelos mesmos juízes que participaram

do teste-piloto, pois já estavam capacitados para o uso do instrumento, que foi aplicado de forma concomitante e independente a uma amostra de 33 pacientes, considerando-se os quinze pacientes já avaliados anteriormente no teste-piloto.

Para a avaliação dos indicadores de 1 a 11 e 15, o juiz leu o descritor que define o padrão de qualidade e assinalou, no instrumento de avaliação, se o aspecto do cuidado observado estava adequado ou inadequado em relação ao padrão. Para os indicadores 12, 13 e 14, com base na explicitação do padrão de qualidade, cada juiz registrou a quantidade de checagens e registros de sinais vitais realizados de forma adequada e a quantidade inadequada.

Quanto ao tratamento dos dados, foi utilizada a estatística Kappa (*k*) para análise das variáveis categóricas – indicadores de 1 a 11 e 15 e o Coeficiente de Correlação Interclasses (CCI) para as variáveis contínuas – indicadores 12, 13 e 14, considerando-se o Intervalo de Confiança (IC) de 95%^(12,18).

O coeficiente de Kappa é uma medida de associação, que mede o grau de concordância para além do que seria esperado pelo acaso, com pesos iguais para as discordâncias⁽¹⁷⁾; o Fleiss Kappa é uma extensão da estatística Kappa para avaliar a concordância entre mais de dois avaliadores, sem aplicação de ponderação⁽¹⁹⁾.

O Coeficiente de Correlação Interclasses ou coeficiente de Reprodutibilidade (R) é uma estimativa da fração da variabilidade total de medidas, devido a variações entre os indivíduos⁽²⁰⁻²¹⁾.

Apesar de não fornecer informações detalhadas sobre a estrutura da concordância e discordância⁽²²⁾ a estatística Kappa é melhor do que os percentuais simples de concordância⁽¹⁷⁾. Autores afirmam que, para a estimação da confiabilidade interobservadores, a estatística Kappa e o Coeficiente de Correlação Interclasses são os métodos mais apropriados⁽⁹⁾.

Paralelamente à estatística Kappa e CCI, também foram calculados os percentuais simples de concordância para se ter uma ideia detalhada da confiabilidade e concordância⁽⁹⁾.

Os valores de Kappa foram interpretados da seguinte forma: menor que 0 como indicativo de *no agreement*; entre 0 e 0,19 *poor agreement*; 0,20 a 0,39 *fair agreement*; 0,40 a 0,59 *moderate agreement*; 0,60 a 0,79 *substantial agreement* e 0,80 a 1,00 *almost perfect agreement*⁽²³⁾. Para o CCI foram utilizados os valores de 0,4 a 0,59 razoável reprodutibilidade; 0,6 a 0,74 boa reprodutibilidade e acima de 0,74 excelente reprodutibilidade⁽²⁴⁾.

Os dados categóricos (estatística *k*) foram tratados utilizando-se o *software* Lee, do Laboratório de Epidemiologia e Estatística do Instituto Dante Pazzanese de Cardiologia – Brasil e os dados contínuos (CCI) pelo *software* BioEstat 5.0. O p-value de <0,05 foi considerado estatisticamente significativo.

O estudo foi conduzido em concordância com todos os preceitos éticos, tendo sido aprovado pela direção da instituição e pelo Comitê de Ética em Pesquisas Envolvendo Seres Humanos da Universidade, sob Parecer nº 126/10, CAAEE nº0113.0.268.000-10. Todos os sujeitos de pesquisa, juízes e pacientes, assinaram o Termo de Consentimento Livre e Esclarecido (TCLE).

Resultados

Os resultados do procedimento de validação de face apontaram para a validade aparente de todos os indicadores. Quanto ao teste-piloto, o parecer dos juízes foi favorável no que tange à adequação do instrumento relacionada aos quesitos compreensão e aplicabilidade. Considerando-se que não houve dúvidas nem sugestões, o instrumento foi considerado apropriado para continuidade do teste de confiabilidade interavaliadores.

A Tabela 1, a seguir, apresenta os resultados da aplicação da estatística Kappa para os indicadores de 1 a 11 e 15, além do percentual de concordância simples para todo o conjunto dos indicadores.

Tabela 1 - Determinação da concordância interavaliadores pela aplicação da estatística Kappa aos indicadores de 1 a 11 e 15. Londrina, PR, Brasil, 2013

Indicador*	N [†]	% concordância	Fleiss k	95% CI	p-valor	Concordância
01	33	100	1,000	0,803-1,000	<0,001	Excelente
02	25	100	1,000	0,835-1,000	<0,001	Excelente
03	22	100	1,000	0,852-1,000	<0,001	Excelente
04	25	97	0,970	0,830-1,000	<0,001	Excelente
05a	10	97	0,956	0,795-1,000	<0,001	Excelente
05b	23	94	0,969	0,829-1,000	<0,001	Excelente
05c	04	100	1,000	0,847-1,000	<0,001	Excelente
06a	10	100	1,000	0,845-1,000	<0,001	Excelente
06b	23	97	0,969	0,829-1,000	<0,001	Excelente
06c	02	100	1,000	0,846-1,000	<0,001	Excelente
07	09	100	1,000	0,837-1,000	<0,001	Excelente
08	03	100	1,000	0,803-1,000	<0,001	Excelente
09	05	100	1,000	0,846-1,000	<0,001	Excelente
10	05	100	1,000	0,803-1,000	<0,001	Excelente
11	05	100	1,000	0,846-1,000	<0,001	Excelente
15	27	100	1,000	0,855-1,000	<0,001	Excelente

*Indicadores 1 – 11; 15

1. Identificação do leito do paciente.
2. Identificação de risco para queda do leito.
3. Identificação de acessos venosos periféricos.
4. Verificação de lesões cutâneas pós-infiltrativas.
5. a – Identificação de equipes para infusão venosa (soro de manutenção).
- b – Identificação de equipes para infusão venosa (soro para diluição de medicação).
- c – Identificação de equipes para infusão venosa (soro – droga 1).
6. a – Identificação de frascos de infusão venosa – rótulo (soro de manutenção).
- b – Identificação de frascos de infusão venosa – rótulo (soro para diluição de medicação).
- c – Identificação de frascos de infusão venosa – rótulo (soro – droga 1).
7. Identificação do controle de velocidade de infusão – escala graduada.
8. Identificação de sondas gástricas – oro e nasogástricas para drenagem.
9. Fixação da sonda vesical de demora.
10. Posicionamento da bolsa coletora de diurese, da sonda vesical de demora.
11. Posicionamento do prolongamento distal de drenagem da bolsa coletora de diurese – *spigot*.
15. Elaboração da prescrição diária e completa pelo enfermeiro.

†Número de pacientes em que se aplica a avaliação do indicador.

Os valores da Tabela 1 demonstram um percentual de concordância simples superior a 80%, considerado adequado para todos os indicadores avaliados⁽¹⁴⁾.

A estatística Kappa variou entre 0,956 a 1,000, ou seja, os indicadores de 1 a 11 e 15 demonstram excelente

grau de concordância, estatisticamente significativo – p-valor <0,001⁽²³⁾.

Os indicadores que não obtiveram concordância total pela aplicação do Fleiss *k* foram: indicador 4 – verificação de lesões cutâneas pós-infiltrativas ($k=0,970$, IC 95%: 0,830-1,000); 5a – identificação de equipes para

infusão venosa (soro de manutenção) ($k=0,956$, IC 95%: 0,795- 1,000); 5b – identificação de equipes para infusão venosa (soro para diluição de medicação) ($k=0,969$, IC 95%: 0,829-1,000) e 6b – identificação de frascos de infusão venosa – rótulo ($k=0,969$, IC 95%: 0,829-1,000).

Os resultados da aplicação do Coeficiente de Correlação Interclasses para os indicadores 12, 13 e 14 e o percentual de concordância simples, para o conjunto dos indicadores, estão apresentados na Tabela 2.

Tabela 2 - Determinação da concordância interavaliadores pela aplicação do Coeficiente de Correlação Interclasses aos indicadores de 12, 13 e 14. Londrina, PR, 2013

Indicador*	N	% concordância	CCI	95% CI	p-valor	Reprodutibilidade†
12 - a	559	99,7	0,992	0,983-0,996	<0,001	Excelente
12 - b	81	98,1	0,980	0,959-0,990	<0,001	Excelente
13 - a	56	98,2	0,957	0,914-0,979	<0,001	Excelente
13 - b	44	97,8	0,951	0,903-0,976	<0,001	Excelente
14 - a	354	99,6	0,969	0,938-0,985	<0,001	Excelente
14 - b	64	97,5	0,859	0,732-0,929	<0,001	Excelente

*Indicadores 12 - 14

12. a - Checagem dos procedimentos na prescrição de enfermagem (adequado).

b - Checagem dos procedimentos na prescrição de enfermagem (inadequado).

13. a - Registro de verificação dos sinais vitais prescritos (adequado).

b - Registro de verificação dos sinais vitais prescritos (inadequado).

14. a - Checagem dos procedimentos de enfermagem na prescrição médica (adequado).

b - Checagem dos procedimentos de enfermagem na prescrição médica (inadequado).

†Sinônimo de confiabilidade.

Para aos indicadores 12, 13 e 14 (Tabela 2), o percentual de concordância simples variou entre 89,4 e 92,5%, que caracteriza concordância adequada⁽¹⁴⁾. Os valores de CCI variaram entre 0,859 e 0,992 e os indicadores avaliados obtiveram reprodutibilidade/concordância excelente e estatisticamente significativa – p-valor <0,001⁽²⁴⁾.

O menor Coeficiente de Correlação Interclasses – o CCI obtido foi de 0,859 (IC 95%: 0,732-0,929) para o indicador 14b – checagem dos procedimentos de enfermagem na prescrição médica (adequado); o maior valor de CCI foi de 0,992 (IC 95%: 0,983-0,996) para o indicador 12 a – checagem dos procedimentos na prescrição de enfermagem (adequado), caracterizando reprodutibilidade/concordância excelente para os indicadores avaliados.

Discussão

A confiabilidade interavaliadores, testada pela estatística Kappa, encontrou valores que caracterizam concordância excelente entre os juízes, no que se refere ao construto, aos descritores e aos critérios de avaliação dos indicadores testados.

Cabe ressaltar que o que está sendo medido não é a validade dos resultados, mas, sim, o grau de erro na medida, que se deve às diferenças no escore verdade^(8,10-11). Sendo assim, é possível afirmar que os indicadores 4, 5-a, 5-b, 6-b, 12-a, 12-b, 13-a, 13-b, 14-a e 14-b, que não obtiveram resultados de Fleiss k e CCI de 1,000, não são 100% precisos. Porém, ressalta-se que, por melhor

que seja um instrumento de medida, os escores obtidos nunca estarão livres de erros⁽¹¹⁾.

Os erros que afetam as medidas são classificados em dois subgrupos: os erros aleatórios, que derivam de fatores que afetam a medição da variável de forma acidental em toda a amostra, adicionando variabilidade aos dados, mas não afetando o desempenho médio para o grupo, e os erros sistemáticos, que são causados por qualquer fator que sistematicamente afete a medição da variável em toda a amostra e que tendem a ter um efeito consistentemente positivo ou negativo, sendo, por vezes, em razão disso, considerados como viés de medição⁽¹¹⁾.

Acredita-se que os fatores que possivelmente tenham afetado a precisão da medida, contribuindo para os erros de medição, sejam fatores pessoais transitórios⁽²⁵⁾, como pressa e fadiga, pois um dos juízes encontrava-se participando do estudo após a jornada diária de trabalho. Além disso, existiu a referência por parte dos dois juízes de dificuldades quanto à legibilidade das escritas referentes às checagens nas prescrições médicas e de enfermagem, o que gerou dúvidas em relação ao atendimento total ao padrão de qualidade adotado.

Os valores obtidos para o k e CCI desses indicadores reproduzem o quanto os resultados da aplicação da medida refletem o resultado verdadeiro⁽¹⁰⁾. Mesmo não tendo alcançado um Fleiss k de 1,000, ainda assim os indicadores 4, 5-a, 5-b, 6-b evidenciaram um grau de concordância classificado como excelente – *almost perfect agreement* – k 0,80 a 1,000⁽²³⁾. Os indicadores 12-a, 12-b, 13-a, 13-b, 14-a e 14-b também

evidenciaram concordância excelente, com valores de CCI acima de 0,74⁽²⁴⁾.

Os resultados encontrados sugerem a fidedignidade interavaliadores dos indicadores e instrumento de medida, indicando sua precisão e a possibilidade de utilização para avaliação da qualidade da assistência de enfermagem.

Cabe ressaltar que confiabilidade e concordância não são propriedades fixas dos instrumentos de medida, mas sim, o produto da interação entre o instrumento/ferramenta, os sujeitos/objetos e o contexto da avaliação⁽⁹⁾.

Para que essa propriedade não seja afetada em contextos diversos, é importante estabelecer o controle sobre as variáveis que interferem no processo de medição. Para tanto, os avaliadores devem ser capacitados quanto ao construto, descritores, índice de conformidade ideal, critérios de avaliação, assim como a padronização do procedimento avaliativo.

Conclusão

Com base no estudo, conclui-se que os quinze indicadores de qualidade da assistência de enfermagem estudados, já validados anteriormente pela estratégia de validação de conteúdo, apresentam excelente confiabilidade e reprodutibilidade entre os itens.

A concordância e confiabilidade dos indicadores, estimadas pela estatística Kappa e Coeficiente de Correlação Interclasse demonstraram a relevância desse instrumento para a prática clínica da avaliação da qualidade da assistência de enfermagem.

A confirmação de sua fidedignidade comprova a contribuição do estudo para acrescentar evidências de que o desenvolvimento de instrumentos válidos e fidedignos para a avaliação da qualidade da assistência de enfermagem é possível e imprescindível para gerenciamento efetivo e eficaz da assistência, pois possibilita a identificação de riscos evitáveis, subsidia o planejamento de ações corretivas e direciona estratégias para o reajuste de metas.

Diante da evidência de fidedignidade dos quinze indicadores, acredita-se que sua utilização em outras instituições de saúde possa aprimorar consideravelmente a gestão de enfermagem e, conseqüentemente, a qualidade da assistência e a segurança dos pacientes.

Métodos de desenvolvimento e validação de sistemas de avaliação são extremamente discutidos e empregados em ciências sociais e comportamentais e sua aplicação, neste estudo, demonstra que seus princípios podem ser adaptados para a criação de instrumentos de avaliação das práticas de enfermagem.

As potenciais limitações do estudo incluem a técnica de amostragem intencional e o tamanho da amostra de juízes. O que determinou a opção por essa técnica foi a inviabilidade em conduzir o estudo caso fossem sorteados enfermeiros no seu turno de trabalho.

A opção pelo número mínimo de três juízes justificase pela dificuldade em avaliar, de forma concomitante e independente, todos os pacientes acomodados em enfermarias de três a seis leitos, o que poderia resultar em tumulto, desconforto e constrangimento, além de atuar como contaminante situacional, favorecendo o erro de medição.

Em relação aos indicadores testados, seguramente não abrangem todos os aspectos importantes da assistência de enfermagem, mas, certamente, abarcam cuidados relevantes para a prevenção de riscos e altamente sensíveis à melhoria com investimentos simples, como estratégias educativas. Infelizmente, a seleção desses quinze indicadores resultou na exclusão de outros aspectos relevantes do cuidado, o que pode ser uma limitação do presente estudo.

Referências

1. Rocha ESB, Trevizan MA. Quality management at a hospital's nursing service. *Rev. Latino-Am. Enfermagem*. 2009;17(2):240-5.
2. Kuwabara CCT, Évora YDM, Oliveira MMB. Risk Management in Technovigilance: construction and Validation of a Medical-Hospital Product Evaluation Instrument. *Rev. Latino-Am. Enfermagem*. 2010;18(5):943-51.
3. D'Innocenzo MANP, Cunha ICKO. O movimento pela qualidade nos serviços de saúde e enfermagem. *Rev Bras Enferm*. [Internet]. 2006. [acesso 23 abr 2013];59(1):84-8. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-71672006000100016&lng=en&nrm=iso&tlng=pt
4. Collier R. The challenges of quantifying quality. *CMAJ*. [Internet]. 2010. [acesso 23 abr 2013];182(5):E250. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2842849/pdf/182e250.pdf>
5. Takahashi AA, Barros ALBL, Michel JLMS, Mariana F. Difficulties and facilities pointed out by nurses of a university hospital when applying the nursing process. *Acta Paul Enferm*. [Internet]. 2008. [acesso 23 abr 2013];21(1):32-8. Disponível em: <http://www.scielo.br/pdf/ape/v21n1/04.pdf>
6. Vieira APM, Kurcgant, P. Quality indicators of the management of human resources in nursing: point of view of registered nurses. *Acta Paul Enferm*. [Internet]. 2010. [acesso 23 abr 2013];23(1):11-5. Disponível em: <http://www.scielo.br/pdf/ape/v23n1/02.pdf>

7. Nakrem S, Vinsnes AG, Harkless GE, Paulsen B, Seim A. Nursing sensitive quality indicators for nursing home care: international review of literature, policy and practice International. *J Nurs Stud*. 2009; 46:848-57.
8. Gillespie BM, Polit DF, Hamlin L, Chaboyer W. Developing a model of competence in the operating theatre: Psychometric validation of the Perceived Perioperative Competence Scale-Revised. *Int J Nurs Stud*. [Internet]. 2012. [acesso 23 abr 2013];49(1):90-101. Disponível em: [http://www.journalofnursingstudies.com/article/S0020-7489\(11\)00301-4/fulltext](http://www.journalofnursingstudies.com/article/S0020-7489(11)00301-4/fulltext)
9. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol*. [Internet]. 2011. [acesso 23 abr 2013];64(1):96-106. Disponível em: <http://www.rygforskning.dk/sites/default/files/files/articles/Kottner%20el%20al%20%202011.pdf>
10. Hora HRM, Monteiro GTR, Arica J. Confiabilidade em questionários para qualidade: um estudo com o Coeficiente Alfa de Cronbach. *Prod Prod*. [Internet]. 2010. [acesso 23 abr 2013];11(2):85-103. Disponível em: <http://seer.ufrgs.br/ProdutoProducao/article/view/9321/8252>
11. Trochim WMK. Research methods: knowledge bases. [Internet]. 2nd. ed. Cincinnati (OH): Atomic Dog Publishing; 2006. [acesso 23 abr 2013]. Disponível em: <http://socialresearchmethods.net/kb/>
12. Wuensch KL. The Intraclass Correlation Coefficient. Karl Wuensch's Statistics Lessons [Internet]. Greenville (USA): East Carolina University - Department of Psychology; [2010]; [atualizada em 20 abr 2013; acesso 23 abr 2013]. Disponível em: <http://core.ecu.edu/psyc/wuenschk/StatsLessons.htm>
13. Vituri DW, Matsuda LM. Content validation of quality indicators for nursing care evaluation. *Rev Esc Enferm USP*. [Internet]. 2009. [acesso 23 abr 2013];43(2):429-37. Disponível em: http://www.scielo.br/pdf/reusp/v43n2/en_a24v43n2.pdf
14. Westmoreland D, Wesorick B, Hanson D, Wyngarden K. Consensual validation of clinical practice model practice guidelines. *J Nurs Care Qual*. 2000;14(4):16-27.
15. Crocker L, Llabre M, Miller MD. The generalizability of content validity ratings. *J Educ Measure*. 1988; 25(4):287-99.
16. Lilford RJ, Mohammed MA, Brauhoultz D, Hofer TP. The measurement of active errors: methodological issues, *Qual Saf Health Care*. [Internet]. 2003. [acesso 30 abr 2012];12 Suppl 2:8-12. Disponível em: http://qualitysafety.bmj.com/content/12/suppl_2/ii8.full.pdf+html
17. Stemler SE. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Prac Assess Res Eval*. [Internet]. 2004. [acesso 15 abr 2012];9(4). Disponível em: <http://pareonline.net/getvn.asp?v=9&n=4>
18. Haley SM, Osberg JS. Kappa Coefficient Calculation Using Multiple Ratings Per Subject: A Special Communication. *Phys Ther*. [Internet]. 1989. [acesso 15 abr 2012];69:970-4. Disponível em: <http://ptjournal.apta.org/content/69/11/970>
19. Chang A. Cohen's and Fleiss's Kappa program Concordance in ordinal data. In: Chang A. StatTools Home Page [Internet]. Queensland: Austrália; 2011. [acesso 20 abr 2012]; Disponível em: http://www.stattools.net/CohenKappa_Pgm.php#Fleiss%27s%20Kappa%20from%20rating%20scores
20. Chang A. Intraclass correlation for parametric data Introduction and explanation. In: Chang A. StatTools Home Page [Internet]. Queensland: Austrália; 2011 [acesso 15 abr 2012]. Disponível em: http://www.stattools.net/ICC_Exp.php
21. Wilson-Genderson M, Broder HL, Phillips C. Concordance between caregiver and child reports of children's oral health-related quality of life. *Community Dent Oral Epidemiol*. 2007;35 Suppl 1:32-40.
22. Perroca MG, Gaidzinki RR. Assessing the interrater reliability of an instrument for classifying patients - kappa quotient. *Rev Esc Enferm USP*. [Internet]. 2003. [acesso 23 abr 2013];37(1):72-80. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0080-62342003000100009
23. Zegers M, Bruijne MC, Wagner C, Groenewegen PP, Wal GVD, Vet HCW. The inter-rater agreement of retrospective assessments of adverse events does not improve with two reviewers per patient record. *J Clin Epidemiol*. 2010;63(1):94-112.
24. Fleiss JL. Reliability of measurement. In: Fleiss JL. *The Design and Analysis of Clinical Experiments*. New York (USA): John Wiley; 1999. p. 1-31.
25. Polit DF, Beck CT. Assessing data quality. In: Polit DF, Beck CT. *Nursing research: principles and methods*. 7th ed. Philadelphia (USA): Lippincott Williams & Wilkins; 2004. p. 413-48.

Recebido: 26.4.2013

Aceito: 3.9.2013