



EMPLOYMENT, LABOUR AND SOCIAL AFFAIRS DIRECTORATE

Health Division

Selecting Indicators for Patient Safety at the Health Systems Level in OECD Countries

This document will be presented under item 5 of the draft agenda for the Patient Safety Subgroup meeting and under item 8 of the draft agenda for the Expert Group meeting. It has been prepared by Patrick Romano.

Health Care Quality Indicators Patient Safety Subgroup Meeting/Health Care Quality Indicators Expert Meeting

To be held 24/25-26th October 2007
UEO

Beginning 9:30 a.m. on the first day

DRAFT

SELECTING INDICATORS FOR PATIENT SAFETY AT THE HEALTH SYSTEMS LEVEL IN
OECD COUNTRIES

SUMMARY OF RECENT US EXPERIENCE

PATRICK S. ROMANO
University of California Davis
Center for Healthcare Policy and Research

September 24, 2007

DISCLAIMER

This report was produced under contract to the Organization for Economic Cooperation and Development (OECD). It is intended for use by OECD staff and participants in the Health Care Quality Indicators project. It contains preliminary data that are still subject to change, so it should not be disseminated outside the OECD without the permission of the author. Selected portions of this report were developed with data collected and/or analyzed under contract with the Agency for Healthcare Research and Quality (AHRQ). The information and opinions expressed herein reflect solely the position of the author(s). Nothing herein should be construed to indicate AHRQ support or endorsement of its contents.

DRAFT

BACKGROUND

The OECD Health Care Quality Indicator (HCQI) Project was started in 2001, with the objective of developing a set of indicators that are based on comparable data and that can be used to raise questions for further investigation of quality differences across countries. The HCQI Project will eventually represent the largest effort to track international health care quality that has ever been undertaken. An Expert Group representing the 23 participating countries, the World Health Organization, the European Commission, and leading research organizations, identified five priority areas for initial development of indicators: cardiac care, diabetes mellitus, mental health, patient safety, and prevention/health promotion together with primary care. An expert panel was convened in each of these priority areas, and tasked with identifying and evaluating potential indicators. The Expert Group agreed to apply three evaluation criteria in this process: (1) importance, based on health impact, policy importance, and susceptibility to influence by the health care system; (2) scientific soundness, based on face validity and content validity; and (3) feasibility, based on data availability and reporting burden.

To identify potential indicators in the area of patient safety, the OECD convened a Patient Safety Panel chaired by Dr. John Millar from the Canadian Institute for Health Information. This panel sought indicators that would cover “the following five core domains of patient safety”: hospital-acquired infections, sentinel events, operative and postoperative complications, obstetrics, and other care-related adverse events. A total of 59 indicators from seven different sources were identified by OECD staff, submitted by the Expert Group, or proposed by members of the Patient Safety Panel itself. Through a structured review process modeled on the RAND Corporation’s modified Delphi method, the Patient Safety Panel converged on a final list of 21 indicators that were deemed suitable for international application (based on both importance and scientific soundness). This list was released on October 2004 in OECD Health Technical Paper No. 18, *Selecting Indicators for Patient Safety at the Health Systems Level in OECD Countries*.

Recognizing the high level of interest in patient safety across its member countries, the OECD formed a Patient Safety Expert Group, which met for the first time in June 2006 in Dublin, Ireland. This conference, hosted by the Irish Department of Health and Children, was convened by the OECD to address three issues: (1) getting patient safety data systems on the policy agenda internationally; (2) developing a work plan for improving patient safety data systems and international comparability of patient safety data; and (3) linking data to action to improve patient safety. Presenters described the data systems in their own countries, and their country-specific efforts to use these data to monitor patient safety and to understand variation across hospitals and regions. From the presentations at this conference and a separate survey on patient safety data availability, it became clear that no international database on patient safety exists, very limited data are immediately comparable across countries, and even when such data are available, other factors inhibit their use for international benchmarking.

To foster progress on these difficult issues, the OECD Secretariat proposed, and the Patient Safety Expert Group enthusiastically endorsed, a specific initiative to adapt hospital administrative data systems for assessing and comparing patient safety across member countries. It was agreed that this initiative would focus on the original set of 21 indicators endorsed by the Patient Safety Panel through its structured review in 2004. Twelve of these indicators have been offered since 2002 by the US Agency for Healthcare Research and Quality (AHRQ) as Patient Safety Indicators (PSIs); six of these twelve (e.g., infection due to medical care, decubitus ulcer, postoperative sepsis, accidental puncture or laceration, transfusion reaction, foreign body left in during procedure) have also been applied to children as AHRQ Pediatric Quality Indicators (PDIs) since 2006. As a result, AHRQ and the users of its Quality Indicators software have accumulated substantial experience with these 12 indicators. This experience may be useful to the OECD’s Patient Safety Expert Group, as it moves forward with implementation of the proposed international study of patient safety.

DRAFT

This report was written to summarize available evidence from the USA regarding the validity of the 12 AHRQ Patient Safety Indicators/Pediatric Quality Indicators endorsed by the OECD Patient Safety Panel in 2004. This evidence derives from a variety of sources and study methodologies, and its applicability to data systems from other countries is uncertain. However, the USA has more experience with hospital administrative data systems than most OECD member countries, and hence problems affecting data in the USA are likely to apply to other countries as well. Exceptions to this general principle will be discussed.

DOMAINS OF VALIDITY

Applying general concepts of validation to the field of patient safety measurement, we can identify several potential domains of validity. Some of these domains are easier to tap than others; unfortunately, the most challenging domains to tap are often the most useful for health care providers who wish to understand the meaning of the data.

Content validity addresses the extent to which the content of a measure is consistent with professional knowledge about health care quality and the outcomes of high-quality care. Consensual validation is the most rigorous approach for assessing the content validity of health care quality indicators, because it requires agreement or near-consensus among professionals from different disciplines, different regions, and different practice environments. Ideally, the expert panels convened for consensual validation represent all of the disciplines involved in treating the condition(s) of interest, include at least 8-10 members, and discuss all of the relevant evidence supporting use of the quality indicator.

Construct validity addresses the extent to which one purported measure of quality is correlated with other measures with which a high correlation would be expected, according to the conceptual framework underlying quality improvement research. The most common application of this approach, known as convergent validity, is to estimate correlations between measures of the process of care and measures of the outcomes of that care. Process measures include both implicit assessments, in which health professionals review available documents or other evidence to formulate a global assessment of quality, and explicit assessments, which focus on specific evidence-based diagnostic tests or treatments. Explicit process measures are typically preferred, because they are often (but not always) based on randomized controlled trials, which are relatively immune to bias from unmeasured confounders. Another approach to construct validation is to study associations between outcome measures and structural indicators, such as nurse staffing levels and skill mix, that have previously been shown to represent markers of quality. Finally, some authors test the construct that any meaningful adverse outcome should be associated with other adverse outcomes. Applied to patient safety measurement, this construct posits that in-hospital adverse events should be associated with subsequent mortality, readmissions, prolonged length of stay, and long-term disability.

Finally, criterion validity addresses the extent to which one purported measure of quality is correlated with other, better measures of the same phenomenon. It implies the existence of a “gold standard” that can be used to evaluate less costly – and presumably less accurate – measurement methods. Applied to patient safety measurement, this approach typically involves comparing indicators based on routinely collected administrative data with indicators of the same outcomes based on more complex linked data, in-depth medical record review, physician/nurse interview, patient interview, or even direct observation. Criterion validity may represent the strongest validation approach, but its applicability is often limited by the lack of an accepted “gold standard.”

OVERVIEW

In the following sections, recently available evidence from the USA about the validity of the 12 PSIs/PDIs is summarized, in each of these three domains of validity. The most recent evidence comes from the AHRQ PSI Validation Pilot project, which was undertaken to: (1) gather evidence on the scientific acceptability of the PSIs; (2) improve guidance on the interpretation and use of the indicators; (3) evaluate potential refinements to the indicator specifications; (4) develop medical record abstraction tools that users can apply to review potentially preventable adverse events and to identify opportunities for improvement; and (5) to develop an infrastructure for conducting validation studies on an ongoing basis. Phase 1 of this study focused on selected infections due to medical care, postoperative pulmonary embolus or deep vein thrombosis, postoperative sepsis, accidental puncture or laceration, and iatrogenic pneumothorax; the first four of these PSIs were endorsed by the OECD Patient Safety Panel. However, data collection for the AHRQ PSI Validation Pilot project just ended as of the date of this report, and analyses of the data have just begun, so presented are only preliminary results in rather general terms. More detailed results will be shared with the OECD as they become available, later in 2007.

It should also be noted that preliminary evidence from the AHRQ PSI Validation Pilot, and supporting evidence from other studies, suggests that several of the PSI's that were not endorsed by the OECD Patient Safety Panel score highly in the domains of validity. Iatrogenic pneumothorax and postoperative wound dehiscence are the most striking examples. If the concerns about translation into ICD-10 (and international procedure coding systems) can be resolved, the OECD may wish to reconsider these indicators.

HOSPITAL-ACQUIRED INFECTIONS SELECTED INFECTIONS DUE TO MEDICAL CARE

OECD rationale and concerns. "Infections related to medical care can be a very serious problem... Many infections acquired in the course of medical care are preventable by proper hygiene, rational use of antibiotics and other measures. The occurrence of nosocomial infection is widely acknowledged to be a valid measure of health care quality... The ICD codes chosen are reasonable but there may be considerable variation in the coding practices."

Content (consensual) validity. Content validity was addressed in the Technical Report accompanying the original release of the AHRQ PSIs. Although panelists' ratings of the "usefulness" of each candidate indicator were used to select the final PSI set, panelists were also asked to rate each indicator on its preventability and its likelihood of being due to medical error. The median rating of this indicator was 7, with indeterminate agreement, on the former dimension, leading to a classification of "acceptable." The median rating was 6, with indeterminate agreement, on the latter dimension, leading to a classification of "unclear."

Construct (convergent, predictive) validity. This indicator rates very highly on predictive validity. Cases from the Nationwide Inpatient Sample (NIS) that were flagged by this PSI had 4.3% excess mortality, 9.6 days of excess hospitalization, and \$38,700 in excess hospital charges, relative to carefully matched controls that were not flagged (Zhan C, Miller M. *JAMA* 2003;290:1868-74). This finding was confirmed in the Veterans Affairs (VA) hospital system, where cases that were flagged by this PSI had 2.7% excess mortality, 4.5-9.5 days of excess hospitalization, and \$7,292-13,816 in excess hospital costs, relative to carefully matched controls that were not flagged (Rivard P, et al. *Med Care Res Rev*; in press). In a study testing construct validity using an implicit process measure of quality (Miller MR, et al., *Am J Med Qual* 2005; 20:239-252), smoothed rates of this PSI among 2,116 hospitals surveyed by the Joint Commission for the Accreditation of Healthcare Organizations (JCAHO) were not associated with summary evaluation scores. Physicians participating in the National Association of Children's Hospitals and Related Institutions' (NACHRI) Pediatric PSI Collaborative reviewed 152 flagged events from 20 hospitals, using

DRAFT

an online tool to assess implicit process of care, and judged 41% to be preventable and only 30% to be clearly non-preventable (Sedman A, et al. *Pediatrics* 2005;115:135-45).

Criterion validity. No evidence about the criterion validity of this indicator was available before its release. We now know that this indicator has a minor problem due to missing data about timing. Two US states have publicly available administrative data that includes a “flag” variable denoting whether each diagnosis was present at admission. The percentage of cases flagged by this PSI for whom the event was reported to be a complication of the hospital stay was 65% in California, 65% in New York, and 60% in the Rochester, Minnesota area (Houchens RL, et al. *Joint Comm J Qual Safety*, in press; Naessens JM, et al. *Med Care* 2007;45:781–8).

Preliminary evidence from the 37 hospitals participating in the AHRQ PSI Validation Pilot Project (N=194) indicates that about 17% of the flagged events were present at admission, and about 22% of cases lacked clear documentation of an infection related to infusion, injection, transfusion, etc., leaving about 61% that were confirmed as iatrogenic complications. The great majority of the confirmed events (circa 80%) were attributable to a venous catheter. Finally, evidence from New York suggests that a significant number of true events may not be ascertained because they occur after hospital discharge; linking 30-day readmissions increased the overall rate of this PSI from 2.02 to 2.52 per 1,000 eligible discharges; 56% of the post-discharge events were complications of hemodialysis access.

Conclusions. Recent evidence on construct validity and criterion validity is moderately supportive of this indicator. A forthcoming change to the ICD-9-CM coding of catheter-associated infections (999.31 = “infection due to central venous catheter”) should further enhance its criterion validity in the USA.

HOSPITAL-ACQUIRED INFECTIONS DECUBITUS ULCER

OECD Rationale. “The occurrence of a decubitus ulcer in a hospitalized patient has a serious negative impact on the individual’s health and often leads to a much prolonged hospital stay... Decubitus ulcers are preventable with good quality nursing care... Thus, the indicator has great clinical plausibility as a patient safety measure. While the indicator is well operationalized, the biggest threat to construct validity is the inability to precisely distinguish between pre-existing and hospital-acquired decubitus ulcers on the basis of administrative data.”

Content (consensual) validity. Content validity was addressed in the Technical Report accompanying the original release of the AHRQ PSIs. Although panelists’ ratings of the “usefulness” of each candidate indicator were used to select the final PSI set, panelists were also asked to rate each indicator on its preventability and its likelihood of being due to medical error. The median rating of this indicator was 8, with agreement, on both dimensions, leading to a classification of “acceptable” on both dimensions.

Construct (convergent, predictive) validity. This indicator rates very highly on predictive validity. Cases from the NIS that were flagged by this PSI had 7.2% excess mortality, 4.0 days of excess hospitalization, and \$10,800 in excess hospital charges, relative to carefully matched controls that were not flagged (Zhan C, Miller M. *JAMA* 2003;290:1868-74). This finding was confirmed in the Veterans Affairs hospital system, where cases that were flagged by this PSI had 6.8% excess mortality, 3.7-5.2 days of excess hospitalization, and \$5,552-6,713 in excess hospital costs, relative to carefully matched controls that were not flagged (Rivard P, et al. *Med Care Res Rev*; in press). In a study testing construct validity using an implicit process measure of quality (Miller MR, et al., *Am J Med Qual* 2005; 20:239-252), smoothed rates of this PSI among 2,116 hospitals surveyed by the JCAHO were not associated with summary evaluation scores. Physicians participating in the NACHRI Pediatric PSI Collaborative reviewed 130 flagged events

DRAFT

from 20 hospitals, using an online tool to assess implicit process of care, and judged 55% to be preventable and only 36% to be clearly non-preventable (Sedman A, et al. *Pediatrics* 2005;115:135-45).

At least two older studies assessed the construct validity of the ICD-9-CM codes mapped to this PSI through correlation with structural measures of nurse staffing. Needleman and Buerhaus found that several measures of nurse staffing were inconsistently associated with the pressure ulcers among medical patients from 799 hospitals in 11 states in 1997, and were independent of pressure ulcers among major surgery patients (Needleman J, et al. *N Engl J Med* 2002;346:1415–22). However, nursing skill mix (RN hours/licensed nurse hours) was significantly associated (in the expected direction) with the pressure ulcer rate among 352 and 295 California hospitals in 1992 and 1994, respectively, and also among 126 and 131 New York hospitals in the same years (Lichtig LK, et al. *J Nurs Adm* 1999;29(2):25-33). Total licensed nurse hours per acuity-adjusted patient day were inconsistently associated with pressure ulcers in that study.

Criterion validity. Several small studies and one large study provided data about the criterion validity of this indicator before its release. Berlowitz et al. (*JAGS* 1999;47:692-696) found that the sensitivity of a discharge diagnosis of pressure ulcer among all patients transferred from VA hospitals to VA nursing homes in 1996 was 31% overall, or 54% for stage IV (deep) ulcers. The overall sensitivity increased modestly since 1992 (26.0%), and was slightly but statistically significantly better among medical patients than among surgical patients (33% versus 26%).

The critical weakness of this indicator is its inability to distinguish pressure ulcers that were present at admission from pressure ulcers that developed during a hospital stay. Two US states have publicly available administrative data that includes a “flag” variable denoting whether each diagnosis was present at admission. The percentage of cases flagged by this PSI for whom the event was reported to be a complication of the hospital stay was only 11% in California, 14% in New York, and 18% in the Rochester, Minnesota area (Houchens RL, et al. *Joint Comm J Qual Safety*, in press; Naessens JM, et al. *Med Care* 2007;45:781–8).

Conclusions. Although the US estimates of criterion validity are suspiciously low, and may reflect hospitals’ efforts to deflect blame, this indicator should be used very cautiously (if at all) for comparing hospital performance without information about the timing of the diagnosis and more recent estimates of sensitivity.

OPERATIVE AND POSTOPERATIVE COMPLICATIONS COMPLICATIONS OF ANESTHESIA

OECD rationale and concerns. “Death due to anesthesia has become rare... By contrast morbid events... are much more prevalent, ranging from postoperative nausea through to equipment failure. Many such events (apart from the obvious ones given above) may be difficult to classify as preventable or avoidable... Further, the studies reviewed to support this indicator have mainly been observational without control group, reducing the face validity of the indicators. The key problem here would seem to be the difficulty in classifying the majority of adverse events as preventable or avoidable. Adequate criteria appear not to be available. There may also be underreporting in administrative data.”

Content (consensual) validity. Content validity was addressed in the Technical Report accompanying the original release of the AHRQ PSIs. Although panelists’ ratings of the “usefulness” of each candidate indicator were used to select the final PSI set, panelists were also asked to rate each indicator on its preventability and its likelihood of being due to medical error. The median rating of this indicator was about 7.5, with agreement, on both dimensions, leading to a classification of “acceptable” on both dimensions.

More recent evidence has raised grave doubts about the content validity of this PSI, which is heavily dependent on External Cause of Injury (E) codes. In fact, of 1,356 cases flagged by this indicator in the 2000 NIS, only 8 would still be flagged if all E-codes were deleted from the definition. This finding is problematic because regulations for use of E-codes vary from state to state; only 16 of the 36 states that contributed to AHRQ's State Inpatient Databases (SID) in 2002 required reporting of E-codes, and many others only have one E-code field. California, South Carolina, and Washington do not require reporting of E870-E879 (including E876.3 for "endotracheal tube wrongly placed during anesthetic procedure," which is in the PSI definition). Empirical analyses have confirmed that the apparent prevalence of this PSI is highly dependent on the number of diagnosis fields used in the analysis, because E-codes are often appended after the full list of other diagnosis codes. Indeed, the prevalence of this PSI decreased more than 20% when the number of diagnosis fields was truncated from 15 to 10 (as many states participating in the SID do). Similarly, this is the only PSI for which prevalence differs markedly between the NIS and Medicare's inpatient claims database, within identical Medicare-eligible age strata, presumably because of diagnosis truncation on Medicare claims (Mathematica Policy Research, Inc. *Medicare Quality Monitoring System (MQMS) Report: Patient Safety, 2000 and 2001 Final Report*. 2003). The unique susceptibility of this indicator to diagnosis truncation was also demonstrated by the Dallas-Fort Worth Hospital Council.

The *ICD-9-CM Official Guidelines for Coding and Reporting*, reviewed with staff from the Cooperating Parties, confirms that the codes mapped to this indicator are not intended to capture patient safety events: "Codes from the E930-E949 series must be used to identify the causative substance for an adverse effect of drug... *correctly prescribed and properly administered*. The effect, such as tachycardia, delirium, gastrointestinal hemorrhaging, vomiting... is coded and followed by the appropriate code from the E930-E949 series. Adverse effects of therapeutic substances correctly prescribed and properly administered (toxicity, synergistic reaction, side effect, and idiosyncratic reaction) may be due to (1) differences among patients... and (2) drug-related factors..."

Construct (convergent, predictive) validity. This indicator rates poorly on predictive validity. Cases from the NIS that were flagged by this PSI had no excess mortality, only 0.2 excess hospital days, and only \$1,600 in excess hospital charges, relative to carefully matched controls that were not flagged (Zhan C, Miller M. *JAMA* 2003;290:1868-74). The reported differences in hospital length-of-stay and total charges were neither meaningful nor statistically significant. In a study testing construct validity using an implicit process measure of quality (Miller MR, et al., *Am J Med Qual* 2005; 20:239-252), smoothed rates of this PSI among 2,116 hospitals surveyed by the JCAHO were not associated with summary evaluation scores. Physicians participating in the NACHRI Pediatric PSI Collaborative reviewed 74 flagged events from 20 hospitals, using an online tool to assess implicit process of care, and judged only 15% to be preventable and 50% to be clearly non-preventable (Sedman A, et al. *Pediatrics* 2005;115:135-45). These non-preventable events included routine side effects of anesthesia and analgesia, such as vomiting, sedation, and pruritis.

Criterion validity. No evidence about the criterion validity of this indicator was available before its release. This indicator does not appear to have a significant problem due to missing data about timing. Two US states have publicly available administrative data that includes a "flag" variable denoting whether each diagnosis was present at admission. The percentage of cases flagged by this PSI for whom the event was reported to be a complication of the hospital stay was 100% in California, 100% in New York, and 6% in the Rochester, Minnesota area, where outpatient surgery may be coded differently (Houchens RL, et al. *Joint Comm J Qual Safety*, in press; Naessens JM, et al. *Med Care* 2007;45:781-8).

DRAFT

Conclusions. Recent evidence on face validity and construct validity suggests that this indicator should not be used to compare hospital performance across US states or nations. Without persuasive evidence of criterion validity, this indicator should be used very cautiously, if at all.

OPERATIVE AND POSTOPERATIVE COMPLICATIONS POSTOPERATIVE HIP FRACTURE

OECD rationale and concerns. “As hip fracture can have devastating consequences including pain, loss of function and, sometimes, death, it has immense clinical significance. When hip fracture occurs in the postoperative period, it can reflect inappropriate prescribing by medical staff (*e.g.*, use of long-acting sedatives) or inadequate nursing procedures (*e.g.*, lack of patient monitoring and bedrail use)... Although it may be impossible to completely eliminate postoperative falls leading to hip fracture, through appropriate prescribing and use of pain relief medication and good nursing care, these should be kept to a minimum. For surgical cases the coding quality has been found to be high, even though there may also be underreporting in administrative data.”

Content (consensual) validity. Content validity was addressed in the Technical Report accompanying the original release of the AHRQ PSIs. Although panelists’ ratings of the “usefulness” of each candidate indicator were used to select the final PSI set, panelists were also asked to rate each indicator on its preventability and its likelihood of being due to medical error. The median rating of this indicator was 8 (with agreement) on the former dimension, and 7 (with indeterminate agreement) on the latter dimension, leading to a classification of “acceptable” on both dimensions.

Construct (convergent, predictive) validity. This indicator rates very highly on predictive validity. Cases from the NIS that were flagged by this PSI had 4.5% excess mortality, 5.2 days of excess hospitalization, and \$13,400 in excess hospital charges, relative to carefully matched controls that were not flagged (Zhan C, Miller M. *JAMA* 2003;290:1868-74). In a study testing construct validity using an implicit process measure of quality (Miller MR, et al., *Am J Med Qual* 2005; 20:239-252), smoothed rates of this PSI among 2,116 hospitals surveyed by the JCAHO were not associated with summary evaluation scores.

Historically, this indicator evolved from one of the “flags” in Iezzoni’s Complications Screening Program (CSP). Explicit process of care failures in the CSP validation study were relatively frequent among cases flagged on this indicator (76% of major surgery patients, 54% of medical patients), after excluding patients who had hip fractures at admission, but unflagged controls were not evaluated on the same criteria (Iezzoni LI, et al. *Int J Qual Health Care* 1999; 11:107-18). Physician reviewers identified potential quality problems in 24% of major surgery patients and 5% of medical patients flagged on this indicator, versus 2% of unflagged controls for each risk group (Weingart SN, et al. *Med Care* 2000;38:796-806).

Criterion validity. The original CSP definition of this PSI, which also included some external cause-of-injury codes, had an adequate confirmation rate among major surgical cases sampled from FY1994 Medicare inpatient claims from California and Connecticut (57% according to coders, 71% according to physicians), but a very poor confirmation rate of 11% (according to both reviewers) among medical cases (Lawthers A, et al. *Med Care* 2000; 38:785-95; Weingart SN, et al. *Med Care* 2000;38:796-806). Based on this finding, AHRQ limited this PSI to surgical cases (defined using DRGs).

The critical weakness of this indicator is its inability to distinguish hip fractures that were present at admission from hip fractures that developed during a hospital stay. Two US states have publicly available administrative data that includes a “flag” variable denoting whether each diagnosis was present at admission. The percentage of cases flagged by this PSI for whom the event was reported to be a complication of the hospital stay was only 21% in California, 26% in New York, and 22% in the

DRAFT

Rochester, Minnesota area (Houchens RL, et al. *Joint Comm J Qual Safety*, in press; Naessens JM, et al. *Med Care* 2007;45:781–8). Although the PSI logic requires that the hip fracture repair occur after an “index” operating room procedure, this logic fails because some hip fracture patients are not surgically repaired and others have comorbid diagnoses (e.g., coronary artery disease) that lead to procedural intervention before the hip fracture is repaired.

Conclusions. Although alternative logic to improve positive predictive value (PPV) is currently being tested, this indicator should be used very cautiously (if at all) for comparing hospital performance without information about the timing of the diagnosis.

OPERATIVE AND POSTOPERATIVE COMPLICATIONS POSTOPERATIVE PE OR DVT

OECD rationale and concerns. “Because PE/DVT can cause unnecessary prolongation of hospital stays as well as unnecessary pain, suffering and death, this indicator has important financial and quality improvement implications. PE/DVT can be prevented through the appropriate use of anticoagulants and other preventive measures. Given the numerous measures undertaken to reduce postoperative PE/DVT, this indicator has clinical plausibility. Coding of those events should be unambiguous, but PE/DVT is known to frequently go undiagnosed. Thus, health systems with better monitoring practices may be mislabeled as having unusually high event rates.”

Content (consensual) validity. Content validity was addressed in the Technical Report accompanying the original release of the AHRQ PSIs. Although panelists’ ratings of the “usefulness” of each candidate indicator were used to select the final PSI set, panelists were also asked to rate each indicator on its preventability and its likelihood of being due to medical error. The median ratings of this indicator on the former dimension from two independent panels were 7 and 6, with indeterminate agreement and disagreement, respectively, leading to classifications of “acceptable” and “unclear,” respectively. The median ratings of this indicator on the latter dimension were 6 and 3, with indeterminate agreement, leading to classifications of “unclear” and “unacceptable,” respectively.

Construct (convergent, predictive) validity. This indicator rates very highly on predictive validity. Cases from the NIS that were flagged by this PSI had 6.6% excess mortality, 5.4 days of excess hospitalization, and \$21,700 in excess hospital charges, relative to carefully matched controls that were not flagged (Zhan C, Miller M. *JAMA* 2003;290:1868-74). This finding was confirmed in the Veterans Affairs hospital system, where cases that were flagged by this PSI had 6.1% excess mortality, 4.5-5.5 days of excess hospitalization, and \$7,205-9,064 in excess hospital costs, relative to carefully matched controls that were not flagged (Rivard P, et al. *Med Care Res Rev*; in press). In a study testing construct validity using an implicit process measure of quality (Miller MR, et al., *Am J Med Qual* 2005; 20:239-252), smoothed rates of this PSI among 2,116 hospitals surveyed by the JCAHO were marginally ($p=0.06$) associated with summary evaluation scores, in the expected direction. In addition, hospitals with high smoothed rates of this PSI were less likely to receive favorable accreditation decisions than hospitals with lower rates. Physicians participating in the NACHRI Pediatric PSI Collaborative reviewed 126 flagged events from 20 hospitals, using an online tool to assess implicit process of care, and judged only 29% to be preventable and 48% to be clearly non-preventable (Sedman A, et al. *Pediatrics* 2005;115:135-45).

Historically, this indicator evolved from one of the “flags” in Iezzoni’s Complications Screening Program (CSP). Explicit process of care failures in the CSP validation study were relatively frequent among cases flagged on this indicator (72% of major surgery patients, 69% of medical patients), after excluding patients who had DVT/PE at admission, but unflagged controls were not evaluated on the same criteria (Iezzoni LI, et al. *Int J Qual Health Care* 1999; 11:107-18). Major surgical cases flagged on this indicator and unflagged controls differed marginally (11% versus 4%, $p=0.09$) on a composite of 17

DRAFT

generic process criteria. Physician reviewers identified potential quality problems in 50% of major surgery patients and 20% of medical patients flagged on this indicator, versus 2% of unflagged controls for each risk group (Weingart SN, et al. *Med Care* 2000;38:796-806).

At least two older studies assessed the construct validity of the ICD-9-CM codes mapped to this PSI through correlation with structural measures of nurse staffing. Needleman and Buerhaus (Needleman J, et al. *N Engl J Med* 2002;346:1415–22) found that nurse staffing was independent of the occurrence of DVT/PE among both major surgical and medical patients from 799 hospitals in 11 states in 1997. However, Kovner and Gergen reported that among 506 community hospitals in the 1993 NIS, having more registered nurse hours and non-RN hours per adjusted patient day were both associated with a lower rate of DVT/PE after major surgery (Kovner C, Gergen PJ. *Image J Nurs Sch* 1998;30:315-21). Nurse staffing was not associated with the rate of DVT/PE after invasive vascular procedures.

Criterion validity. The original CSP definition of this PSI, which differed slightly from the current AHRQ definition, had an adequate confirmation rate among major surgical cases sampled from FY1994 Medicare inpatient claims from California and Connecticut (59% according to coders, 70% according to physicians, 68% according to nurses who relied on physician documentation), but a very poor confirmation rate of 28–32% among medical cases (Lawthers A, et al. *Med Care* 2000; 38:785-95; Weingart SN, et al. *Med Care* 2000;38:796-806; McCarthy EP, et al. *Med Care* 2000;38:868-76). Several smaller, older studies also suggested adequate sensitivity and PPV of PE codes among surgical patients, although the sensitivity of DVT codes was notably poorer (Keeler E, et al. *Assessing quality of care for hospitalized Medicare patients with hip fracture using coded diagnoses from the Medicare Provider Analysis and Review File* 1991; Romano PS, et al. *Am J Med Qual* 2002;17:145-154; Hawker GA, et al. *J Clin Epidemiol* 1997;50:265-73; Best W, et al. *J Am Coll Surg* 2002;194:257-66). Based on these findings, AHRQ limited this PSI to surgical cases (defined using DRGs).

One weakness of this indicator is its inability to distinguish thromboses that were present at admission from thromboses that developed during a hospital stay. Two US states have publicly available administrative data that includes a “flag” variable denoting whether each diagnosis was present at admission. The percentage of cases flagged by this PSI for whom the event was reported to be a complication of the hospital stay was only 46% in California, 43% in New York, and 40% in the Rochester, Minnesota area (Houchens RL, et al. *Joint Comm J Qual Safety*, in press; Naessens JM, et al. *Med Care* 2007;45:781–8). These estimates are suspiciously low, and may reflect hospitals’ misinterpretation of preoperative sonographic findings (e.g., chronic thromboses).

Preliminary analyses of data from the AHRQ PSI Validation Pilot project suggest that the actual percentage of events flagged by this indicator that were present at admission may be 20% or less, but up to 24% of the flagged events may involve upper extremity veins or superficial veins (which are not the target for prevention). Comparing hospital administrative data from the Department of Veterans Affairs against the National Surgical Quality Improvement Program’s clinically abstracted data from 2001, we (Romano et al. *HSR*, in press) recently reported a sensitivity of 56%, PPV of 22%, and positive likelihood ratio of 65. Most of the false positives appear to be attributable to chronic thromboses that were present at admission, upper extremity thromboses, or superficial lower extremity thromboses that did not require anticoagulation.

Finally, evidence from New York suggests that a significant number of true events may not be ascertained because they occur after hospital discharge; linking 30-day readmissions increased the overall rate of this PSI from 9.3 to 11.3 per 1,000; 45% of the post-discharge events were pulmonary emboli.

DRAFT

Conclusions. Although alternative logic to improve PPV is currently being tested, this indicator should be used very cautiously for comparing hospital performance unless validated information is available about the timing of the diagnosis and/or the specific veins involved.

OPERATIVE AND POSTOPERATIVE COMPLICATIONS POSTOPERATIVE SEPSIS

OECD rationale and concerns. “The occurrence of sepsis following surgery is a severe complication with a mortality rate of up to 30%. Even less severe cases will require prolonged ICU treatment for organ failure... Many cases of postoperative sepsis can be prevented through the appropriate use of prophylactic antibiotics, good surgical site preparation, careful and sterile surgical techniques and good post-op care. Sepsis after elective surgery is considered a severe complication. It usually results from less severe infective complications, such as urinary tract infections, pneumonia and wound infection, which should be avoided and/or properly treated. Consequently, this indicator is a plausible patient safety measure. Given the dramatic nature of this complication, it is usually reliably coded in administrative data sources.”

Content (consensual) validity. Content validity was addressed in the Technical Report accompanying the original release of the AHRQ PSIs. Although panelists’ ratings of the “usefulness” of each candidate indicator were used to select the final PSI set, panelists were also asked to rate each indicator on its preventability and its likelihood of being due to medical error. The median rating of this indicator was 6.5 (with agreement) on the former dimension, and 6 (with indeterminate agreement) on the latter dimension, leading to a classification of “unclear” on both dimensions.

Construct (convergent, predictive) validity. This indicator rates very highly on predictive validity. Cases from the NIS that were flagged by this PSI had 21.9% excess mortality, 10.9 days of excess hospitalization, and \$57,700 in excess hospital charges, relative to carefully matched controls that were not flagged (Zhan C, Miller M. *JAMA* 2003;290:1868-74). This finding was confirmed in the Veterans Affairs hospital system, where cases that were flagged by this PSI had 30.2% excess mortality, 5.7-18.8 days of excess hospitalization, and \$13,395-31,262 in excess hospital costs, relative to carefully matched controls that were not flagged (Rivard P, et al. *Med Care Res Rev*; in press). In a study testing construct validity using an implicit process measure of quality (Miller MR, et al., *Am J Med Qual* 2005; 20:239-252), smoothed rates of this PSI among 2,116 hospitals surveyed by the JCAHO were marginally ($p=0.10$) associated with summary evaluation scores, in the expected direction.

At least one older study assessed the construct validity of the ICD-9-CM codes mapped to this PSI through correlation with structural measures of nurse staffing. Needleman and Buerhaus (Needleman J, et al. *N Engl J Med* 2002;346:1415-22) found that nurse staffing was independent of the occurrence of sepsis among both major surgical and medical patients from 799 hospitals in 11 states in 1997.

Criterion validity. Several small studies provided limited data about the criterion validity of the ICD-9-CM codes mapped to this indicator before its release. Unfortunately, several of these studies either did not clearly document their ICD-9 definitions (Massanari RM, et al. *Am J Public Health* 1987;77:561-4; Belio-Blasco C, et al. *Infect Control Hosp Epidemiol* 2000;21:24-7) or did not stratify the subgroup of patients with a secondary diagnosis of DVT/PE (Barbour GL. *Am J Med Qual* 1993;8:2-5). In comparison with the VA’s National Surgical Quality Improvement Program database from 123 hospitals in 1994-95, in which “systemic sepsis” was defined by a positive blood culture with systemic manifestations of sepsis within 30 days after surgery, ICD-9-CM diagnoses had a sensitivity of 37% and a PPV of 30% (Best W, et al. *J Am Coll Surg* 2002;194:257-66).

This indicator has a minor problem due to missing data about timing. Two US states have publicly available administrative data that includes a “flag” variable denoting whether each diagnosis was present

DRAFT

at admission. The percentage of cases flagged by this PSI for whom the event was reported to be a complication of the hospital stay was 73% in California, 70% in New York, and 76% in the Rochester, Minnesota area (Houchens RL, et al. *Joint Comm J Qual Safety*, in press; Naessens JM, et al. *Med Care* 2007;45:781–8).

Preliminary evidence from the 33 hospitals participating in the AHRQ PSI Validation Pilot Project (N=137) indicates that about 17% of the flagged events were present at admission, and about 16% of cases lacked clear documentation of sepsis, bacteremia, or SIRS with infection, leaving about 67% that were confirmed as complications. (However, an additional 18% of flagged cases may have been ineligible because the reviewer perceived the “index” surgery as being non-elective.) The primary site of infection was catheter-related in about 31%, other bloodstream in about 6%, lungs in about 29%, surgical site in about 16%, and urinary tract in about 8%. Comparing hospital administrative data from the Department of Veterans Affairs against the National Surgical Quality Improvement Program’s clinically abstracted data from 2001, we (Romano et al. *HSR*, in press) recently reported a sensitivity of 37%, PPV of 45%, and positive likelihood ratio of 131. Most of the “false positives” appear to be patients with clinical evidence of sepsis, who were treated for presumptive sepsis, but lacked “definitive evidence of infection.”

Conclusions. Recent evidence on construct validity and criterion validity is moderately supportive of this indicator, but raises questions about sensitivity.

OPERATIVE AND POSTOPERATIVE COMPLICATIONS

ACCIDENTAL PUNCTURE OR LACERATION (formerly “technical difficulty with procedure”)

OECD rationale and concerns. “While accidental cut, puncture, perforation or laceration during a surgical procedure is a recognized risk, for example of abdominal surgery, elevated rates of such complications may indicate systems problems, such as inadequate surgical training or fatigued surgeons... Traditionally such adverse events were dealt with by peer review procedures, the effectiveness of which in reducing future frequency of adverse events has not been proven. It remains to be seen whether national schemes such as those already referred to will eventually demonstrate more convincing effects. There has been considerable dispute over what to include and not to include in this measure... No convincing evidence on validity is available from previous studies.”

Content (consensual) validity. Content validity was addressed in the Technical Report accompanying the original release of the AHRQ PSIs. Although panelists’ ratings of the “usefulness” of each candidate indicator were used to select the final PSI set, panelists were also asked to rate each indicator on its preventability and its likelihood of being due to medical error. The median rating of this indicator was 7, with agreement, on the former dimension, leading to a classification of “acceptable.” The median rating was 6, with indeterminate agreement, on the latter dimension, leading to a classification of “unclear.”

Construct (convergent, predictive) validity. This indicator rates highly on predictive validity. Cases from the NIS that were flagged by this PSI had 2.2% excess mortality, 1.3 days of excess hospitalization, and \$8,300 in excess hospital charges, relative to carefully matched controls that were not flagged (Zhan C, Miller M. *JAMA* 2003;290:1868-74). This finding was confirmed in the Veterans Affairs hospital system, where cases that were flagged by this PSI had 3.2% excess mortality, 1.4-3.1 days of excess hospitalization, and \$3,359-6,880 in excess hospital costs, relative to carefully matched controls that were not flagged (Rivard P, et al. *Med Care Res Rev*; in press). In a study testing construct validity using an implicit process measure of quality (Miller MR, et al., *Am J Med Qual* 2005; 20:239-252), smoothed rates of this PSI among 2,116 hospitals surveyed by the JCAHO were significantly ($p < 0.01$) associated with summary evaluation scores, in the expected direction. Physicians participating in the NACHRI Pediatric PSI Collaborative reviewed 133 flagged events from 20 hospitals, using an online tool to assess implicit

DRAFT

process of care, and judged 65% to be preventable and only 14% to be clearly non-preventable (Sedman A, et al. *Pediatrics* 2005; 115:135-45).

Criterion validity. Several studies that were published before the release of this indicator offered conflicting conclusions about the criterion validity of the underlying ICD-9 codes. For example, a study of laparoscopic cholecystectomy in 18 Ontario hospitals in 1991-95 (Taylor B. *CMAJ* 1998;158:481-5) found that 95% (99/104) of patients with an ICD-9 code of 998.2 or E870.0 had a confirmed injury to the bile duct or gallbladder (although only 27% were “clinically significant”). A similar study of all cholecystectomies performed in Western Australia between 1988 and 1994 reported that these two codes had a sensitivity of 40% (19/48) and a PPV of 23% (19/84) in identifying bile duct injuries (Valinsky LJ, et al. *J Clin Epidemiol* 1999;52:893-901). Among 185 total knee replacement patients from 5 Ontario hospitals in 1984-90, Hawker et al. (*J Clin Epidemiol* 1997;50:265-73) found that the sensitivity and PPV of codes describing “miscellaneous mishaps during or as a direct result of surgery” were 86% (6/7) and 55% (6/11), respectively. Romano et al. (*Am J Med Qual* 2002;17:145-154) identified 19 of 45 chart-confirmed episodes of accidental puncture or laceration using discharge abstracts of discectomy patients at 30 California hospitals in 1990-91, with only one false positive.

This indicator does not appear to have a significant problem due to missing data about timing. Two US states have publicly available administrative data that includes a “flag” variable denoting whether each diagnosis was present at admission. The percentage of cases flagged by this PSI for whom the event was reported to be a complication of the hospital stay was 87% in California, 87% in New York, and 85% in the Rochester, Minnesota area (Houchens RL, et al. *Joint Comm J Qual Safety*, in press; Naessens JM, et al. *Med Care* 2007;45:781-8). Preliminary evidence from the 43 hospitals participating in the AHRQ PSI Validation Pilot Project (N=230) indicates that about 5% of the flagged events were present at admission, and about 7% of cases lacked clear documentation of an accidental puncture or laceration, leaving about 88% that were confirmed as complications. Of these events, about 80% occurred in the abdomen or pelvis, and about 15% in the chest. Confirming the importance of the injury, about 59% were repaired as part of the same “index” surgery, and about 10% were repaired during a subsequent return to the operating room.

We have very limited evidence about the sensitivity of this indicator. Investigators in New York systematically searched their hospital administrative data for procedure codes suggesting repair of iatrogenic injuries, and reported that this PSI may have missed 27% of bladder injuries from hysterectomy, 21% of bowel injuries from cholecystectomy, 47% of abdominal injuries from lysis of adhesions, 54% of abdominal injuries from nephroureterectomy, and 20% of spinal injuries from lumbar surgery. The AHRQ Support for Quality Indicators team is currently evaluating whether these procedure codes can be added to the AHRQ PSI definition to improve its sensitivity, without significantly compromising its PPV.

Conclusions. Recent evidence on construct validity and criterion validity is moderately supportive of this indicator. Forthcoming changes in the indicator logic, to incorporate selected procedure codes, should further enhance its criterion validity in the USA.

SENTINEL EVENTS

TRANSFUSION REACTION

OECD rationale and concerns. “The risk of adverse outcome from erroneous transfusion rivals or exceeds current estimates of the risk of acquiring infectious disease by transfusion (Linden et al., 2000)... The use of systems designed to prevent specific errors may be helpful (such as convenient access to standard operating procedures instructions in work areas, a blood component lock system that will not allow the access of a component unless there is patient wristband and blood component match, etc.).

DRAFT

Recent studies on human error in medicine followed methods derived from the experience gained while analyzing large-scale technological disasters (Eagle et al., 1992; Reason, 1990). They recognized that medical, like technological, accidents nearly always require the conjunction of two types of failures: active failures, mistakes happening while performing a task, and latent failures, or management system errors...”

Content (consensual) validity. Content validity was addressed in the Technical Report accompanying the original release of the AHRQ PSIs. Although panelists’ ratings of the “usefulness” of each candidate indicator were used to select the final PSI set, panelists were also asked to rate each indicator on its preventability and its likelihood of being due to medical error. The median ratings of this indicator on the former dimension from two independent panels were 7 and 8, with disagreement and indeterminate agreement, respectively, leading to classifications of “unclear” from both panels. The median ratings of this indicator on the latter dimension were 7 and 5.3, with indeterminate agreement and disagreement, respectively, leading to classifications of “acceptable” and “unclear,” respectively.

Construct (convergent, predictive) validity. This indicator rates moderately on predictive validity. Cases from the NIS that were flagged by this PSI had no excess mortality, but they did have 3.4 days of excess hospitalization and \$18,900 in excess hospital charges, relative to carefully matched controls that were not flagged (Zhan C, Miller M. *JAMA* 2003;290:1868-74). However, these differences were not statistically significant, presumably due to the extreme rarity of this event.

Criterion validity. No evidence about the criterion validity of this indicator was available before its release. This indicator may have a minor problem due to missing data about timing. Two US states have publicly available administrative data that includes a “flag” variable denoting whether each diagnosis was present at admission. The percentage of cases flagged by this PSI for whom the event was reported to be a complication of the hospital stay was 58% in California and 78% in New York (Houchens RL, et al. *Joint Comm J Qual Safety*, in press). Due to the extreme rarity of this event, no other information about criterion validity is available at this time.

Conclusions. Recent evidence on construct validity and criterion validity is moderately supportive of this indicator.

SENTINEL EVENTS

FOREIGN BODY LEFT IN DURING PROCEDURE

OECD rationale and concerns. “Errors relating to the failure to remove surgical instruments at the end of a procedure... are no less common than the better known mishaps such as wrong-site surgery. However, many cases of retained foreign body do not cause harm, although some clearly do. Therefore JCAHO sentinel event policy specifically mentions that “unintentionally retained foreign body without major permanent loss of function” does not require reporting.

Although surgeons and operating room teams rely on the practice of sponge, sharp and instrument counts as a means to eliminate retained foreign bodies, practices are not standardized... even single events may signal a serious system failure that should be addressed... There is only one known study demonstrating indirect evidence of the effectiveness of sponge and instrument counts. There are hints that process redesign in surgical procedures could lead to improvement for example errors in sponge counts are attributed to team fatigue, difficult operations, sponges “sticking together” or staff accepting apparently incompatible counts without re-checking... the event seems a clinically plausible indicator of system failure. Without sufficient research evidence, it is difficult to judge whether this particular construct has specific problems. In a general sense... it may suffer from underreporting.”

DRAFT

Content (consensual) validity. Content validity was addressed in the Technical Report accompanying the original release of the AHRQ PSIs. Although panelists' ratings of the "usefulness" of each candidate indicator were used to select the final PSI set, panelists were also asked to rate each indicator on its preventability and its likelihood of being due to medical error. The median ratings of this indicator on the former dimension from two independent panels were 8 and 7.5, with agreement, leading to classifications of "acceptable" from both panels. The median ratings of this indicator on the latter dimension were 8 and 7, with agreement and indeterminate agreement, respectively, leading to classifications of "acceptable" from both panels.

Construct (convergent, predictive) validity. This indicator rates highly on predictive validity. Cases from the NIS that were flagged by this PSI had 2.1% excess mortality, 2.1 days of excess hospitalization, and \$13,300 in excess hospital charges, relative to carefully matched controls that were not flagged (Zhan C, Miller M. *JAMA* 2003;290:1868-74). In a study testing construct validity using an implicit process measure of quality (Miller MR, et al., *Am J Med Qual* 2005; 20:239-252), smoothed rates of this PSI among 2,116 hospitals surveyed by the JCAHO were not associated with summary evaluation scores. Physicians participating in the NACHRI Pediatric PSI Collaborative reviewed 49 flagged events from 20 hospitals, using an online tool to assess implicit process of care, and judged 51% to be preventable and only 29% to be clearly non-preventable (Sedman A, et al. *Pediatrics* 2005;115:135-45).

Criterion validity. No evidence about the criterion validity of this indicator was available before its release. This indicator may have a problem due to missing data about timing. Two US states have publicly available administrative data that includes a "flag" variable denoting whether each diagnosis was present at admission. The percentage of cases flagged by this PSI for whom the event was reported to be a complication of the hospital stay was 64% in California, 76% in New York, and 54% in the Rochester, Minnesota area (Houchens RL, et al. *Joint Comm J Qual Safety*, in press; Naessens JM, et al. *Med Care* 2007;45:781-8). Due to the rarity of this event, no other information about criterion validity is available at this time.

Conclusions. Recent evidence on construct validity and criterion validity is moderately supportive of this indicator.

OBSTETRICS

BIRTH TRAUMA – INJURY TO NEONATE

OECD rationale and concerns. "Birth trauma can lead to prolonged disability of the infant requiring substantial resources for rehabilitation and care. Birth trauma injury is preventable. Occurrence of mortality or morbidity in childbirth may be due to system failure, poor antenatal treatment, or poor obstetric practice. This indicator has been widely used in the obstetric community, although it is most commonly based on chart review rather than administrative data... The indicator appears to be well operationalized. However, it may be necessary to exclude or adjust for additional high-risk conditions to ensure comparability of this indicator across countries."

Content (consensual) validity. Content validity was addressed in the Technical Report accompanying the original release of the AHRQ PSIs. Although panelists' ratings of the "usefulness" of each candidate indicator were used to select the final PSI set, panelists were also asked to rate each indicator on its preventability and its likelihood of being due to medical error. The median rating of this indicator was 7, with indeterminate agreement, on the former dimension, leading to a classification of "acceptable." The median rating was 6, with disagreement, on the latter dimension, leading to a classification of "unclear."

Construct (convergent, predictive) validity. This indicator rates poorly on predictive validity. Cases from the NIS that were flagged by this PSI had no excess mortality, no excess hospital days, and only \$300 in

DRAFT

excess hospital charges, relative to carefully matched controls that were not flagged (Zhan C, Miller M. *JAMA* 2003;290:1868-74). The reported difference in hospital charges was neither meaningful nor statistically significant. In a study testing construct validity using an implicit process measure of quality (Miller MR, et al., *Am J Med Qual* 2005; 20:239-252), smoothed rates of this PSI among 2,116 hospitals surveyed by the JCAHO were not associated with summary evaluation scores.

Criterion validity. One significant study provided data about the criterion validity of this indicator before its release. Among 669 newborns at Georgetown University Hospital who had a discharge diagnosis of birth trauma, only 25% had sustained a significant injury to the head, neck, or shoulder (Hughes C, et al. *Arch Otolaryngol Head Neck Surg* 1999;125:193-9). This indicator is not likely to have a significant problem due to missing data about timing, because the indicator is inherently limited to in-hospital births.

Conclusions. Recent evidence on construct validity and criterion validity is inconclusive regarding this indicator, but significant questions have been raised.

OBSTETRICS

OBSTETRIC TRAUMA – VAGINAL DELIVERY (now divided according to with/without instrumentation; e.g., forceps or vacuum)

OECD rationale and concerns. “Third and fourth degree perineal laceration can produce significant long term morbidity of women undergoing childbirth... Obstetric trauma during delivery is often preventable. The percentage of deliveries involving third and fourth degree lacerations is a useful quality indicator of obstetrical care and can assist in reducing the morbidity from extensive perineal tears. The indicator appears to be well operationalized. However, it may be necessary to exclude or adjust for additional high-risk conditions to ensure comparability of this indicator across countries.”

Content (consensual) validity. Content validity was addressed in the Technical Report accompanying the original release of the AHRQ PSIs. Although panelists’ ratings of the “usefulness” of each candidate indicator were used to select the final PSI set, panelists were also asked to rate each indicator on its preventability and its likelihood of being due to medical error. The median rating of this indicator was 7, with agreement, on the former dimension, leading to a classification of “acceptable.” The median rating was 5, with disagreement, on the latter dimension, leading to a classification of “unclear.”

Construct (convergent, predictive) validity. This indicator rates moderately on predictive validity. Cases from the NIS that were flagged by this PSI had no excess mortality, but they did have 0.05-0.07 excess hospital days and up to \$220 in excess hospital charges, relative to carefully matched controls that were not flagged (Zhan C, Miller M. *JAMA* 2003;290:1868-74). The reported differences in hospital length-of-stay and total charges were small, but statistically significant given the large number of events. In a study testing construct validity using an implicit process measure of quality (Miller MR, et al., *Am J Med Qual* 2005; 20:239-252), smoothed rates of this PSI among 2,116 hospitals surveyed by the JCAHO were positively (counterintuitively) associated ($p=0.04$) with summary evaluation scores, but only in the subset of women with forceps or vacuum deliveries.

Criterion validity. No evidence about the criterion validity of this indicator was available before its release. This indicator is not likely to have a significant problem due to missing data about timing, because the indicator is inherently limited to women who have an in-hospital delivery. The best data about criterion validity come from the California Obstetric Validation Study (Romano PS, et al. *Obstet Gynecol* 2005;106(4):717-725), which involved a stratified random cluster sample of 1,662 records from 52 hospitals (51% vaginal), of which over 97% were reviewed by an “expert” coder and obstetric nurse abstractor. This PSI demonstrated a sensitivity of 90% and a PPV of 90-95%; adjusting for the complex stratified sampling design increased the sensitivity to 93% but decreased the PPV to 73%. A subsequent

DRAFT

study based on a clinical research data set with 393 positive (3rd/4th degree tears) and 383 negative vaginal deliveries (Brubaker L, et al. *Obstet Gynecol* 2007;109(5):1141-5) reported a sensitivity of 77% and a specificity of 99.7% for this indicator. PPV could not be estimated due to the sampling design, but should be approximately 93% given a typical prevalence of 5%.

Conclusions. Recent evidence on construct validity of this indicator is inconclusive, but the evidence on criterion validity is quite supportive.

OBSTETRICS

OBSTETRIC TRAUMA – CESAREAN DELIVERY

OECD rationale and concerns. See previous indicator.

Content (consensual) validity. Content validity was addressed in the Technical Report accompanying the original release of the AHRQ PSIs. Although panelists' ratings of the "usefulness" of each candidate indicator were used to select the final PSI set, panelists were also asked to rate each indicator on its preventability and its likelihood of being due to medical error. The median rating of this indicator was 7, with agreement, on the former dimension, leading to a classification of "acceptable." The median rating was 5, with disagreement, on the latter dimension, leading to a classification of "unclear."

Construct (convergent, predictive) validity. This indicator rates moderately on predictive validity. Cases from the NIS that were flagged by this PSI had no excess mortality, but they did have 0.4 excess hospital days and \$2,700 in excess hospital charges, relative to carefully matched controls that were not flagged (Zhan C, Miller M. *JAMA* 2003;290:1868-74). The reported differences in hospital length-of-stay and total charges were small, but statistically significant given the large number of events. In a study testing construct validity using an implicit process measure of quality (Miller MR, et al., *Am J Med Qual* 2005; 20:239-252), smoothed rates of this PSI among 2,116 hospitals surveyed by the JCAHO were not associated with summary evaluation scores.

Criterion validity. No evidence about the criterion validity of this indicator was available before its release. This indicator is not likely to have a significant problem due to missing data about timing, because the indicator is inherently limited to women who have an in-hospital delivery. The best data about criterion validity come from the California Obstetric Validation Study, which involved a stratified random cluster sample of 1,662 records from 52 hospitals (30% primary cesarean, 19% repeat cesarean, 51% vaginal)

Conclusions. Recent evidence on construct validity of this indicator is inconclusive, but the evidence on criterion validity is quite supportive.