

Episiotomy (Overuse)

Section 1. Basic Measure Information

1.A. Measure Name

Episiotomy (Overuse)

1.B. Measure Number

0083

1.C. Measure Description

Please provide a non-technical description of the measure that conveys what it measures to a broad audience.

This measure is aimed at preventing inappropriate and unnecessary episiotomies during birth, which can lead to adverse outcomes.

Denominator statement: Percentage of patients, regardless of age, who gave birth vaginally (without shoulder dystocia), during a 12-month period who underwent an episiotomy.

This measure was developed by the American Medical Association (AMA)-convened Physician Consortium for Performance Improvement[®] (PCPI), which is a key member of the Pediatric Measurement Center of Excellence (PMCoE) consortium. The PMCoE is funded by the Agency for Healthcare Research and Quality (AHRQ) and includes the following consortium members: the American Academy of Pediatrics; American Board of Pediatrics; American Board of Medical Specialties; Northwestern University; Truven Health Analytics (formerly Thomson Reuters); Children's Hospital and Health System, Milwaukee; Medical College of Wisconsin; and the American Medical Association (AMA).

1.D. Measure Owner

AMA-convened Physician Consortium for Performance Improvement[®] (PCPI[™]) is the measure owner. The AMA has copyright on the measure set.

1.E. National Quality Forum (NQF) ID (if applicable)

N/A

1.F. Measure Hierarchy

Please note here if the measure is part of a measure hierarchy or is part of a measure group or composite measure. The following definitions are used by AHRQ's National Quality Measures Clearinghouse and are available at <http://www.qualitymeasures.ahrq.gov/about/hierarchy.aspx>:

1. Please identify the name of the collection of measures to which the measure belongs (if applicable). A collection is the highest possible level of the measure hierarchy. A collection may contain one or more sets, subsets, composites, and/or individual measures.

None.

2. Please identify the name of the measure set to which the measure belongs (if applicable). A set is the second level of the hierarchy. A set may include one or more subsets, composites, and/or individual measures.

Prenatal/Perinatal Performance Measurement Set

3. Please identify the name of the subset to which the measure belongs (if applicable). A subset is the third level of the hierarchy. A subset may include one or more composites, and/or individual measures.

None.

4. Please identify the name of the composite measure to which the measure belongs (if applicable). A composite is a measure with a score that is an aggregate of scores from other measures. A composite may include one or more other composites and/or individual measures. Composites may comprise component measures that can or cannot be used on their own.

None.

1.G. Numerator Statement

Patients who underwent an episiotomy.

1.H. Numerator Exclusions

None.

1.I. Denominator Statement

All patients, regardless of age, who gave birth vaginally (without shoulder dystocia), during a 12-month period.

1.J. Denominator Exclusions

Babies with shoulder dystocia are excluded from the measure.

1.K. Data Sources

Check all the data sources for which the measure is specified and tested.

Electronic medical record.

If other, please list all other data sources in the field below.

Section 2: Detailed Measure Specifications

Provide sufficient detail to describe how a measure would be calculated from the recommended data sources, uploading a separate document (+ Upload attachment) or a link to a URL. Examples of detailed measure specifications can be found in the CHIPRA Initial Core Set Technical Specifications Manual 2011 published by the Centers for Medicare & Medicaid Services. Although submission of formal programming code or algorithms that demonstrate how a measure would be calculated from a query of an appropriate electronic data source are not requested at this time, the availability of these resources may be a factor in determining whether a measure can be recommended for use.

Please see attachments for full eSpecifications and coding spreadsheets. Below is an overview of our technical specifications process.

The PMCoE Center of Excellence adopted the PCPI specification process, which places emphasis on developing comprehensive measure specifications for electronic health records (EHRs) so they can provide relevant clinical data on patients and actionable feedback to providers. There are several data sources available for collecting performance measures; generally, different data sources require different sets of measure specifications due to the structure of the systems storing the data. The PCPI recognizes that EHRs are the state-of-the-art for clinical encounters and is focusing significant resources and expertise toward specifying and testing measures to be included in EHRs, as they hold the promise of providing the relevant clinical data for measures and for providing feedback to physicians and other health care providers that is timely and actionable.

The type of specifications developed for this measurement set are aligned with the PCPI approach to focus on the development of EHR specifications for new measure development projects. While the PCPI values prospective claims reporting programs and the data these programs can provide, the PCPI is looking to leverage the data in EHRs. This new focus will

align the PCPI with national initiatives that highlight the benefits and wealth of data that EHRs bring to health care.

The measure specifications for this submission include the following components:

1. A text description of the measure.
2. The Data Requirements Table, which outlines the data elements that are required for the measure, including the identification of the clinical vocabularies applicable to a given data element, the NQF Quality Data Model category and State, as well as the timing parameters for each data element.
3. A visual flow diagram that uses Boolean logic to identify the Initial Patient Population, Exclusions, Denominator, Numerator, and Exceptions included in the measure.
4. Measure calculation.
5. Value sets for each of the data elements.

The measure specification provides the required information to collect the data needed to calculate the quality measure. The AMA-PCPI, through PMCoE, will make full measure specifications for the measure available for public use in accordance with the terms detailed in the Notice of Grant Award. Please see the attached written statement from AMA-PCPI and PMCoE.

Section 3. Importance of the Measure

In the following sections, provide brief descriptions of how the measure meets one or more of the following criteria for measure importance (general importance, importance to Medicaid and/or CHIP, complements or enhances an existing measure). Include references related to specific points made in your narrative (not a free-form listing of citations).

3.A. Evidence for General Importance of the Measure

Provide evidence for all applicable aspects of general importance:

- **Addresses a known or suspected quality gap and/or disparity in quality (e.g., addresses a socioeconomic disparity, a racial/ethnic disparity, a disparity for Children with Special Health Care Needs (CSHCN), a disparity for limited English proficient (LEP) populations).**
- **Potential for quality improvement (i.e., there are effective approaches to reducing the quality gap or disparity in quality).**
- **Prevalence of condition among children under age 21 and/or among pregnant women**

- **Severity of condition and burden of condition on children, family, and society (unrelated to cost)**
- **Fiscal burden of measure focus (e.g., clinical condition) on patients, families, public and private payers, or society more generally, currently and over the life span of the child.**
- **Association of measure topic with children’s future health – for example, a measure addressing childhood obesity may have implications for the subsequent development of cardiovascular diseases.**
- **The extent to which the measure is applicable to changes across developmental stages (e.g., infancy, early childhood, middle childhood, adolescence, young adulthood).**

This measure was developed by the AMA-PCPI, which is a key member of the PMCoE consortium. The AMA-convened Physician Consortium for Performance Improvement[®] (PCPI[™]) is a national, physician-led initiative dedicated to improving patient health and safety through the identification and development of evidence-based clinical performance measures and measurement resources that enhance the quality of patient care and foster accountability. The PCPI is nationally recognized for measure development, specification and testing of measures, and enabling use of measures in EHRs. The PCPI’s measure development resources include a measure testing protocol, a position statement on the evidence base required for measure development, a composite framework, specification and categorization of measure exceptions, and an outcomes measure framework. The PCPI is made up of over 170 member organizations and individuals, including national medical specialty societies, State medical societies, health care professional organizations, Federal agencies, individual members, and other groups interested in improving the quality of health care. Today the PCPI portfolio includes measures in more than 46 clinical areas with over 280 individual measures.

The current quality gap in the number of episiotomies that are preformed has been a focus for quality improvement in recent years. Improving the inappropriate rate of episiotomies will have implications for mothers, babies, and providers, such as pediatricians.

Numerous studies have shown that restricted use of episiotomy results in considerable reduction in maternal morbidity due to perineal lacerations, as well as additional complications that may arise in the months and years after giving birth. The American College of Obstetricians and Gynecologists (ACOG) developed an evidence-based guideline on episiotomy. Mediolateral episiotomy is associated with difficulty of repair, greater blood loss, and possibly, increased early postpartum discomfort. Median episiotomy is associated with a greater risk for extension to include the anal sphincter or rectum. Reported complications of episiotomy include bleeding, infection, abscess formation, and dehiscence (ACOG, 2006).

The rate of episiotomy varies among individual clinicians, practice sites, and health systems. Moreover, there may be disparities in the rate of episiotomy among minority patients. A 2005 study that appeared in the Journal of American Medical Association showed there are no increased or better outcomes with the use of episiotomies, and they sometimes result in more

harm than good (Hartmann, 2005). Evidence does not support maternal benefits traditionally ascribed to routine episiotomy. In fact, outcomes with episiotomy can be considered worse since some proportion of women who would have had lesser injury instead had a surgical incision. This study also concluded that routine episiotomies increased:

- The need for stitching.
- Experience of pain and tenderness.
- The healing period.
- The likelihood of leaking stool or gas.
- Pain with intercourse.

Adopting this measure would help to close the quality gaps in episiotomy use, and patients in the Medicaid population would benefit from improved postpartum outcomes and reduced maternal injury during childbirth. These safety and quality issues illustrate the need for this measure in the CHIPRA program.

References

American College of Obstetricians and Gynecologists. ACOG Practice Bulletin. Episiotomy. Clinical Management Guidelines for Obstetrician-Gynecologists. *Obstet Gynecol* 2006 Apr; 107(4):957-62.

Hartmann K, Viswanathan M, Palmieri R, et al. Outcomes of routine episiotomy: a systematic review. *JAMA* 2005; 293(17):2141-8.

3.B. Evidence for Importance of the Measure to Medicaid and/or CHIP

Comment on any specific features of this measure important to Medicaid and/or CHIP that are in addition to the evidence of importance described above, including the following:

- **The extent to which the measure is understood to be sensitive to changes in Medicaid or CHIP (e.g., policy changes, quality improvement strategies).**
- **Relevance to the Early and Periodic Screening, Diagnostic and Treatment benefit in Medicaid (EPSDT).**
- **Any other specific relevance to Medicaid/CHIP (please specify).**

This measure would fill a gap in the Medicaid and CHIP programs' core set of children's health care quality measures aimed at providing services and treatment to promote healthy birth and prevent premature birth. The measure will provide a mechanism to help assess the appropriateness of deliveries and prevent adverse neonatal outcomes. This measure is of

particular importance for CHIPRA in that it is high impact with Medicaid patients and addresses concerns related to both mother and baby.

We encourage the use of this measure by physicians, other health care professionals, and health care systems or health plans, where appropriate. This clinical performance measure is designed for practitioner- and/or system-level quality improvement to achieve better outcomes for maternity care patients and their babies.

3.C. Relationship to Other Measures (if any)

Describe, if known, how this measure complements or improves on an existing measure in this topic area for the child or adult population, or if it is intended to fill a specific gap in an existing measure category or topic. For example, the proposed measure may enhance an existing measure in the initial core set, it may lower the age range for an existing adult-focused measure, or it may fill a gap in measurement (e.g., for asthma care quality, inpatient care measures).

This measure is harmonized with the National Quality Forum (NQF)-endorsed episiotomy measure developed by Christiana Care Health Services in conjunction with the National Perinatal Information Center/Quality Analytic Services (NPIC/QAS). The aforementioned measure is specified at the facility-level; this measure is specified at the individual provider-level.

Section 4. Measure Categories

CHIPRA legislation requires that measures in the initial and improved core set, taken together, cover all settings, services, and topics of health care relevant to children. Moreover, the legislation requires the core set to address the needs of children across all ages, including services to promote healthy birth. Regardless of the eventual use of the measure, we are interested in knowing all settings, services, measure topics, and populations that this measure addresses. These categories are not exclusive of one another, so please indicate "Yes" to all that apply.

Does the measure address this category?

- a. Care Setting – ambulatory: No.**
- b. Care Setting – inpatient: Yes.**
- c. Care Setting – other – please specify: No.**
- d. Service – preventive health, including services to promote healthy birth: No.**
- e. Service – care for acute conditions : No.**
- f. Service – care for children with acute conditions: No.**
- g. Service – other (please specify): No.**
- h. Measure Topic – duration of enrollment: Yes.**
- i. Measure Topic – clinical quality: Yes.**
- j. Measure Topic – patient safety: No.**
- k. Measure Topic – family experience with care: No.**

- l. **Measure Topic – care in the most integrated setting:** No.
- m. **Measure Topic other (please specify):** No.
- n. **Population – pregnant women:** Yes.
- o. **Population – neonates (28 days after birth) (specify age range):** No.
- p. **Population – infants (29 days to 1 year) (specify age range):** No.
- q. **Population – pre-school age children (1 year through 5 years) (specify age range):** No.
- r. **Population – school-aged children (6 years through 10 years) (specify age range):** No.
- s. **Population – adolescents (11 years through 20 years) (specify age range):** No.
- t. **Population – other (specify age range):** No.
- u. **Other category (please specify):** No.

Section 5. Evidence or Other Justification for the Focus of the Measure

The evidence base for the focus of the measures will be made explicit and transparent as part of the public release of CHIPRA deliberations; thus, it is critical for submitters to specify the scientific evidence or other basis for the focus of the measure in the following sections.

5.A. Research Evidence

Research evidence should include a brief description of the evidence base for valid relationship(s) among the structure, process, and/or outcome of health care that is the focus of the measure. For example, evidence exists for the relationship between immunizing a child or adolescent (process of care) and improved outcomes for the child and the public. If sufficient evidence existed for the use of immunization registries in practice or at the State level and the provision of immunizations to children and adolescents, such evidence would support the focus of a measure on immunization registries (a structural measure).

Describe the nature of the evidence, including study design, and provide relevant citations for statements made. Evidence may include rigorous systematic reviews of research literature and high-quality research studies.

The evidence behind the inappropriate use of episiotomy and its link to an increased risk of adverse outcomes for mothers and babies is well established. Research evidence includes clinical practice guidelines, numerous research studies on the use of episiotomies (including inappropriate use), and systematic evidence reviews that highlight the risks associated with episiotomy.

Clinical Evidence Base Available for Measure:

Evidence-based clinical practice guidelines that were reviewed for this project:

- American College of Obstetricians and Gynecologists.
- American Academy of Family Physicians.
- Centers for Disease Control.
- United States Preventive Services Task Force.
- Veterans Administration/Department of Defense Clinical Practice Guideline for Pregnancy Management.
- Society of Obstetricians and Gynecologists of Canada.

The measure is supported by ACOG guidelines and reviews of literature that report on systematic evidence reviews of the use of episiotomies.

ACOG Episiotomy Clinical Recommendations (2006):

- Restricted use of episiotomy is preferable to routine use of episiotomy (Level A).
- Median episiotomy is associated with higher rates of injury to the anal sphincter and rectum than is mediolateral episiotomy (Level A).
- Mediolateral episiotomy may be preferable to median episiotomy in selected cases (Level B).
- Routine episiotomy does not prevent pelvic floor damage leading to incontinence (Level B).

Several published research papers reported the results of systematic evidence reviews on the use of episiotomy:

1. A report published by the Agency for Healthcare Research and Quality (AHRQ), *The Use of Episiotomy in Obstetrical Care: A Systematic Review*, was evaluated:

Discussion/findings by key question:

Key Question 1: Episiotomy and maternal postpartum outcomes trials of fair to poor quality provide consistent findings that clearly support limited use of episiotomy. Routine episiotomy achieves none of the short-term goals that it has been hypothesized to achieve. Indeed, routine use is harmful to the degree that it creates an unnecessary surgical incision as compared to less severe posterior perineal trauma in women for whom episiotomy has not been performed.

Key Question 2: Episiotomy incision type and maternal morbidity. A single study found that women with midline episiotomy had a significantly greater rate of anal sphincter injuries than women with mediolateral episiotomy. Fifteen treatment groups did not report differences in pain

or satisfaction with intercourse at 3 months. Because of considerable methodological flaws in this trial (poor internal validity), any conclusions must be drawn cautiously. However, because differences in sphincter injury rates are clinically important, we consider the finding of increased risk of severe injury with midline episiotomy compared to mediolateral episiotomy to be relevant observational evidence.

Key Question 3: Repair of perineal defect and maternal morbidity. Limited but consistent evidence favored two-layer repair over three-layer repair; limited and inconsistent evidence favored continuous over interrupted sutures. Evidence was insufficient to comment on comparisons between standard and rapidly absorbed sutures, tissue adhesive and absorbable sutures, or non-absorbable and absorbable sutures. We found no evidence that treated catgut is superior to untreated catgut with regard to perineal morbidity; the former may in fact be associated with higher morbidity. The evidence suggests short-term advantages for avoidance of repeat perineal tear associated with the use of polyglycolic-acid sutures compared to chromic-catgut sutures. Three major classes of suture material (non-absorbable, absorbable, and tissue adhesive) and two subtypes of sutures (treated versus untreated and standard versus rapidly absorbed) were studied, all in the presence of different approaches to the method of suturing. Thus, the individual effects of the materials themselves cannot be examined. Likewise, methods of repair were examined in the context of different materials, both among and within studies for different stages of repair. We are unable to assess the true effects of a certain method of repair because we cannot tell whether outcomes are confounded or modified by suture material.

Key Question 4: Episiotomy and urinary incontinence, fecal incontinence, and pelvic floor defects. The prospective studies examined did not identify improvements in continence for urine or stool or in pelvic floor muscle function among women who had an episiotomy compared to those who had not. This finding includes comparison to women who had spontaneous lacerations of similar severity. Several authors reported decrements in pelvic floor function among women who had an episiotomy. Only a single study, using multivariable models, found that episiotomy was an independent predictor of urinary continence. In the majority of other studies using multivariate models—adjusting for factors such as parity, neonatal weight, and length of second-stage labor—episiotomy was not an independent risk factor for incontinence. Taken in total, this literature, predominantly of fair to poor quality, does not support use of episiotomy for the purpose of preventing pelvic floor defects, urinary incontinence, or incontinence of stool or flatus. These studies are limited because they do not follow women long enough to detect disease occurrence. At present, the assumption that intermediate variables, such as pelvic muscle strength measured by perineometry, urodynamic test results, or early reports of symptoms, can predict later disease has not been validated. Prospective evaluation only during the months after birth when the pelvic floor is still in a recovery and stabilization period may be misleading. Conclusions about whether episiotomy prevents or increases risk for incontinence and prolapse later in adult life cannot be reached from currently available randomized and cohort studies.

Key Question 5. Episiotomy and future sexual function. The studies addressing this question need to be considered in two groups: mediolateral episiotomy and median episiotomy. From the clinical trials of episiotomy strategy—liberal versus restrictive—one trial addressed each type of incision, and one directly compared the two incision types. None found substantive differences in sexual function. The preponderance of the studies, however, supported a conclusion that

degree of perineal trauma is associated with probability of pain with intercourse in a dose-response fashion, such that greater perineal injury is associated with greater probability of pain. Measures that are more complex than those typically used in this literature are needed to understand properly the relationships between perineal trauma and future sexual function. Specific factors such as prior sexual function and current libido, in addition to factors such as duration of second-stage labor, size of infant, and lactation status, need to be incorporated into multivariable models to derive more informative and less biased estimates of the long-term effects of episiotomy or to determine that they do not exist.

2. A JAMA article described a systematic evidence review of the best evidence available on maternal outcomes of routine vs. restrictive use of episiotomy.

Evidence Synthesis:

Fair to good evidence from clinical trials suggests that immediate maternal outcomes of routine episiotomy—including severity of perineal laceration, pain, and pain medication use—are not better than those with restrictive use of episiotomy. Evidence is insufficient to provide guidance on choice of midline vs. mediolateral episiotomy. Evidence regarding long-term sequelae is fair to poor. Incontinence and pelvic floor outcomes have not been followed long enough to reach the age range in which women are most likely to have sequelae. With this caveat, relevant studies are consistent in demonstrating no benefit from episiotomy for prevention of fecal and urinary incontinence or pelvic floor relaxation. Likewise, no evidence suggests that episiotomy reduces impaired sexual function; pain with intercourse was more common among women with episiotomy.

Conclusions:

Evidence does not support maternal benefits traditionally ascribed to routine episiotomy. In fact, outcomes with episiotomy can be considered worse, since some proportion of women who would have had lesser injury instead had a surgical incision. Episiotomy is among the most common surgical procedures experienced by women in the United States; 30 to 35 percent of vaginal births include episiotomy. Episiotomy became routine practice well before emphasis on using outcomes research to inform practice. In seeking to establish an evidence base to support or refute the use of episiotomy, randomized clinical trials in the middle to late 1980s found that routine episiotomy compared with restrictive use was associated with higher risk of anal sphincter and rectal injuries and precluded a woman from giving birth with an intact or minimally damaged perineum. Larger trials in more varied populations followed in the 1990s, with similar results. Investigators also began to assess longer-term outcome, such as persistent pain, pelvic floor defects, urinary and rectal continence, and sexual function and satisfaction.

Despite decades of research, which many interpret as definitive evidence against routine use of episiotomy, little professional consensus has developed about the appropriateness of routine use. Lack of consensus is illustrated by variation in use. At 18 Philadelphia hospitals studied in the mid-1990s, 42 percent of women overall had an episiotomy, while hospital averages ranged from 20 to 73 percent. From 1987 to 1992, Low and colleagues (2000) documented clinician-level variation from 13.3 to 84.6 percent, with an average of 51 percent among spontaneous term births in a prospectively enrolled population of uncomplicated births. Wide variation existed

among both midwives and physicians. Variation has also been reported by time of day and by facility type, size, and location. Obstetric health care practitioners who view episiotomy favorably endorse survey items that state that episiotomy should be used to “prevent perineal trauma and to prevent pelvic floor relaxation and the consequences of pelvic floor relaxation, such as bladder prolapse and urinary incontinence.” Furthermore, they agree with the statement that they “prefer to employ episiotomy frequently because it is easier to repair than the laceration that occurs when episiotomy is not used.” Simultaneous belief in prevention of future sequelae and ease of repair creates potential for misattributed motivations.

National data on use of episiotomy show a consistent decline over the prior two decades. However, persistent wide practice variation suggests that episiotomy use is heavily driven by local professional norms, experiences in training, and individual practitioner preference rather than variation in the needs of individual women at the time of vaginal birth. Our goal was to refocus attention on routine episiotomy by systematically reviewing the best evidence available about the maternal outcomes of routine vs. restrictive use of episiotomy, including type of episiotomy. Specifically, we sought to describe maternal outcomes such as degree of perineal injury and pain close to the time of birth, as well as longer-term outcomes such as urinary and fecal incontinence, pelvic floor defects, and sexual dysfunction.

References

American College of Obstetricians and Gynecologists (ACOG). ACOG Practice Bulletin. Episiotomy. Clinical Management Guidelines for Obstetrician-Gynecologists. *Obstet Gynecol* 2006 Apr; 107(4):957-62.

Hartmann K, Viswanathan M, Palmieri R, et al. Outcomes of routine episiotomy: a systematic review. *JAMA* 2005; 293(17):2141-8.

Low LK, Seng JS, Murtland TL, et al. Clinician-specific episiotomy rates: impact on perineal outcomes. *J Midwifery Womens Health* 2000 Mar-Apr; 45(2):87-93.

Viswanathan M, Hartmann K, Palmieri R, et al. The use of episiotomy in obstetrical care: a systematic review. Summary, Evidence Report/Technology Assessment No. 112. (Prepared by the RTI-UNC Evidence-based Practice Center, under Contract No. 290-02-0016.) AHRQ Pub. No. 05-E009-1. Rockville, MD: Agency for Healthcare Research and Quality; May 2005.

5.B. Clinical or Other Rationale Supporting the Focus of the Measure (optional)

Provide documentation of the clinical or other rationale for the focus of this measure, including citations as appropriate and available.

Section 6. Scientific Soundness of the Measure

Explain the methods used to determine the scientific soundness of the measure itself. Include results of all tests of validity and reliability, including description(s) of the study sample(s) and methods used to arrive at the results. Note how characteristics of other data systems, data sources, or eligible populations may affect reliability and validity.

6.A. Reliability

Reliability of the measure is the extent to which the measure results are reproducible when conditions remain the same. The method for establishing the reliability of a measure will depend on the type of measure, data source, and other factors.

Explain your rationale for selecting the methods you have chosen, show how you used the methods chosen, and provide information on the results (e.g., the Kappa statistic). Provide appropriate citations to justify methods.

Analytic Method

The study sample for reliability testing is being derived from the urban, tertiary-care hospital with an EHR system integrating inpatient and outpatient data. The EHR system is certified for the Medicare and Medicaid EHR Incentive Programs. Data being used in the analysis are from a patient population of 12,108 for 2010. We are carrying out an assessment of measure reliability applying a reliability coefficient in the form of the signal to noise ratio (SNR). In SNR analysis, reliability is the measure of confidence in differentiating performance between physicians or other providers.¹ The signal is the variability in measured performance that can be explained by real differences in physician performance, and the noise is the total variability in measured performance. Reliability is then the ratio of the physician-to-physician variance to the sum of the physician-to-physician variance plus the error variance specific to a physician:

Reliability = Variance (physician-to-physician) / [Variance (physician-to-physician) + Variance (physician-specific-error)]

Reliability equal to zero implies that all the variability in a measure is attributable to measurement error. Reliability equal to 1.0 implies that all the variability is attributable to real differences in physician performance. Reliability of 0.70 is generally considered a minimum threshold for reliability, and 0.80 is considered very good reliability.²

The SNR reliability testing is being performed using a beta-binomial model. The beta-binomial model assumes the physician performance score is a binomial random variable conditional on the physician's true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates.

Reliability can be estimated at different points. The convention is to estimate reliability at two points: one, at a minimum number of quality reporting events per physician and two, at the

average number of quality reporting events per physician. We set the minimum number required as 10 events. Limiting the reliability analysis to only those physicians with a minimum number of events reduces the bias introduced by the inclusion of physicians without a significant number of events. Reliability testing results from SNR analysis have been included in support of AMA-PCPI measures submitted for NQF endorsement.³

A second phase of reliability testing on the measure also is ongoing at the same sites where feasibility testing was conducted. This approach utilizes parallel forms reliability where measure data elements and performance from an automated report from the EHR are compared to those data from a manual review of the HER—that is, comparison to the gold standard. (See Measure Testing Protocol for PCPI Performance Measures, ama-assn.org/resources/doc/cqi/pcpi-testing-protocol.pdf.)

Reliability Testing Results

A total of 128 physicians had all the required data elements and met the minimum number of quality reporting events (10) for inclusion in the reliability analysis. The average number of quality reporting events for physicians included is 63.34 for a total of 8,236 events. The range of quality reporting events for physicians included is 10 to 159. This measure has high reliability at the average number of quality events and moderate reliability when evaluated at the minimum number of quality reporting events. Reliability at the average number of quality reporting events and at the minimum number of quality reporting events were 0.77 and 0.34, respectively.

Data analyses were conducted using SAS/STAT software, version 8.2 (SAS Institute, Cary, NC).

References

1. Adams JL, Mehrotra A, McGlynn EA. Estimating reliability and misclassification in physician profiling. Santa Monica, CA: RAND Corporation; 2010. www.rand.org/pubs/technical_reports/TR863; Adams JL, Mehrotra A, Thomas JW, et al. Physician cost profiling—reliability and risk of misclassification. *N Engl J Med*. 2010 Mar 18; 62(11):1014-21. <http://www.nejm.org/doi/pdf/10.1056/NEJMsa0906323>; and Scholle SH, Roski J, Adams JL, et al. Benchmarking physician performance: reliability of individual and composite measures. *Am J Manag Care* 2008;14:833-8. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2667340/pdf/nihms-99203.pdf>.
2. Nunnally J, Bernstein I. Psychometric theory. 3rd ed. New York, NY: McGraw-Hill; 1994.
3. National Quality Forum. NQF removes time-limited endorsement for 21 measures: measures now have endorsed status. Press release; January 18, 2012. http://www.qualityforum.org/News_And_Resources/Press_Releases/2012/NQF_Removes_Time-Limited_Endorsement_for_21_Measures; Measures_Now_Have_Endorsed_Status.aspx.

6.B. Validity

Validity of the measure is the extent to which the measure meaningfully represents the concept being evaluated. The method for establishing the validity of a measure will depend on the type of measure, data source, and other factors.

Explain your rationale for selecting the methods you have chosen, show how you used the methods chosen, and provide information on the results (e.g., R2 for concurrent validity).

The measure was assessed for content validity and face validity. Evidence of content validity is provided by looking for agreement among subject matter experts. The performance measure was assessed for content validity by a panel of expert workgroup members during the development process. This expert panel comprised 24 members, with representation from measure methodologists, patient advocacy groups, and the following clinical specialties: anesthesiology, family practice, geriatric medicine, maternal fetal medicine, neonatology, nurse midwife, obstetrics and gynecology, and perinatal nursing. Additional input on the content validity of draft measures was obtained through a 30-day public comment period and by also soliciting comments from a panel of consumer, purchaser, and patient representatives convened by the PCPI specifically for this purpose. All comments received were reviewed by the expert workgroup, and the measures were adjusted as needed.

The expert panel members also assessed the measure's face validity through an online survey. The survey introduction provided the following definition of face validity: "Face validity is the extent to which an empirical measurement appears to reflect that which it is supposed to 'at face value.'" Face validity of an individual measure poses the question of how well the definition and specifications of an individual measure appear to capture the single aspect of care or health care quality as intended. The expert panel was asked to rate their agreement with the following statement: The scores obtained from the measure as specified will accurately differentiate quality across providers. A 5-point Likert scale was used in the survey (1=Strongly Disagree; 2=Disagree; 3=Neither Disagree nor Agree; 4 = Agree 5=Strongly Agree).

The survey results show that for the episiotomy measure, the mean score was 4.46; 100 percent (13/13) of respondents agree or strongly agree that the scores obtained from the measure as specified will accurately differentiate quality across providers; no respondents disagree or strongly disagree that the scores obtained from the measure as specified will accurately differentiate quality across providers.

Section 7. Identification of Disparities

CHIPRA requires that quality measures be able to identify disparities by race, ethnicity, socioeconomic status, and special health care needs. Thus, we strongly encourage nominators to have tested measures in diverse populations. Such testing provides evidence for assessing measure's performance for disparities identification. In the sections below, describe the results of efforts to demonstrate the capacity of this measure to produce results that can be stratified by the characteristics noted and retain the scientific soundness (reliability and validity) within and across the relevant subgroups.

7.A. Race/Ethnicity

We include race and ethnicity as Supplemental Data Elements to collect for each measure to allow for the stratification of measure results by these variables to assess disparities and initiate subsequent quality improvement activities.

The Centers for Disease Control and Prevention (CDC) value sets for race and ethnicity are referenced in the measure specifications to collect race and ethnicity information, which is the requirement for race and ethnicity outlined in the Centers for Medicare & Medicaid Services (CMS) Blueprint.

Also see Section 8.B.1 and Section 8.B.2

7.B. Special Health Care Needs

Not applicable for this measure.

7.C. Socioeconomic Status

We include payer as a Supplemental Data Element to collect for each measure to allow for the stratification of measure results by this variable to assess disparities and initiate subsequent quality improvement activities.

The Payment Typology value set is referenced in the measure specifications to collect payer information, which is the requirement for payer outlined the CMS Blueprint.

Also see Section 8.B.1 and Section 8.B.2

7.D. Rurality/Urbanicity

Future measure testing and implementation will collect data on the location of the patient and provider populations in order to stratify performance and test for variation by location.

7.E. Limited English Proficiency (LEP) Populations

We include preferred language as a Supplemental Data Element to collect for each measure to allow for the stratification of measure results by this variable to assess disparities and initiate subsequent quality improvement activities addressing identified disparities, consistent with recent national efforts to standardize the collection of preferred language data.

The CDC value set is referenced in the measure specifications to collect preferred language information, which is the requirement for preferred language outlined in the CMS Blueprint.

Also see Section 8.B.1 and Section 8.B.2.

Section 8. Feasibility

Feasibility is the extent to which the data required for the measure are readily available, retrievable without undue burden, and can be implemented for performance measurement. Using the following sections, explain the methods used to determine the feasibility of implementing the measure.

8.A. Data Availability

1. What is the availability of data in existing data systems? How readily are the data available?

Data Element Table Tool

The PMCoE Center of Excellence adopted the AMA-PCPI testing methodology which uses the Data Element Table (DET) tool to assess the availability of the data and the technical feasibility and implementation feasibility of the measures. The DET is an Excel workbook designed to capture information that will determine whether or not each site can feasibly collect the data for the measures. It is structured to collect meta-data about each data element necessary to construct each measure stored in the EHR. It will also collect information related to integrity and validity of data collection. Specifically, the DET is designed to capture the following information:

- *Data element information:* Whether or not the data element is captured in the EHR, the data source application, primary user interface data location, data type, coding system, unit of measure, frequency of collection, and calculability within the measure context.
- *Measure integrity information:* An assessment by the testing site as to what degree the measure, as specified, retains the originally stated intention of the measure.
- *Measure validity information:* An assessment by the testing site as to what degree the scores obtained from the measure, as specified, will accurately differentiate quality performance across providers.

The DETs collected responses used to assess technical and implementation feasibility for each measure. Measure technical feasibility was defined as “Can my EHR do this?” and measure implementation feasibility was defined as “Will workflow be used consistently?” The responses were captured in the form of a rating using the following responses:

- “Feasible. Can do today.”
- “Feasible with workflow mod/changes to EHR.”
- “Non-feasible. Unable to do today.”

This information was entered from drop-down options pertaining to the specific criteria and in free text fields for questions related to specific workflow and EHR configurations. The free text

fields and specific narrative questions provide qualitative feedback from the sites, which can be factored into the overall feasibility grade for the measure.

The DET is completed by staff at each testing site. After the completion of the DET by the testing sites, a determination can be made as to which of the measures are relevant for each site. For some sites, all of the measures in the Perinatal/Prenatal Measurement Set may be collected, while for others it may be only a few.

Once the completed DETs were submitted by the test sites, the PMCoE project team conducted quality assurance (QA) of the DETs to ensure the data were complete and ready for analysis. A series of analyses were subsequently performed in order to characterize the feasibility, integrity, and face validity of the measures being tested.

Feasibility testing was conducted at three sites: an urban, tertiary care hospital; an urban, public hospital; and a suburban community hospital.

All three test sites reported that all three of the data elements can be captured in their EHR.

2. If data are not available in existing data systems or would be better collected from future data systems, what is the potential for modifying current data systems or creating new data systems to enhance the feasibility of the measure and facilitate implementation?

Measure Technical Feasibility and Implementation Feasibility

The measure technical feasibility assessment determined how many of the total measure data elements are feasible data elements. A “feasible data element” is one that can be captured by the test site EHR system. The sites assessed technical feasibility for the measure based on the following rating scale:

- “Feasible. Can do today.”
- “Feasible with workflow mod/changes to EHR.”
- “Nonfeasible. Unable to do today.”

The sites also used this scale to assess measure implementation feasibility. Implementation feasibility represents the site’s ability to implement the measure using current workflows and EHRs and addresses issues of projected data reliability related to the consistency with which providers document and capture the data elements needed to implement the measure.

The technical feasibility and implementation feasibility were rated the same for each of the measures. For example, if the technical feasibility of a measure was rated as “Feasible. Can do today,” its implementation feasibility was also rated as “Feasible. Can do today.”

Two of the three sites rated the feasibility of the measure as “Feasible. Can do today.” These two sites capture all of the data elements and can calculate the measure with their current technical configurations. A third site rated the feasibility of the measure as “Feasible with workflow

mod/changes to EHR.” This site captures all three of the measure data elements, however the data are found in free text results, which the site reported as a limitation in this case. Some workflow modifications or changes to this site’s EHR system may be necessary in order to calculate the measure.

8.B. Lessons from Use of the Measure

1. Describe the extent to which the measure has been used or is in use, including the types of settings in which it has been used, and purposes for which it has been used.

Although development of the measure was completed in 2012, there has been limited opportunity to have the measure adopted and implemented. Feasibility and reliability testing of the measures have been conducted in EHRs in a variety of settings, including an urban, tertiary care hospital; an urban, public hospital; and a suburban community hospital; and provide a description of data collection methods and insights into lessons learned. See results presented in Section 6.A. Reliability and Section 8. Feasibility.

2. If the measure has been used or is in use, what methods, if any, have already been used to collect data for this measure?

N/A

3. What lessons are available from the current or prior use of the measure?

N/A

Section 9. Levels of Aggregation

CHIPRA states that data used in quality measures must be collected and reported in a standard format that permits comparison (at minimum) at State, health plan, and provider levels. Use the following table to provide information about this measure’s use for reporting at the levels of aggregation in the table.

For the purpose of this section, please refer to the definitions for provider, practice site, medical group, and network in the Glossary of Terms.

If there is no information about whether the measure could be meaningfully reported at a specific level of aggregation, please write "Not available" in the text field before progressing to the next section.

Level of aggregation (Unit) for reporting on the quality of care for children covered by Medicaid/ CHIP:

State level Can compare States

Intended use: Is measure intended to support meaningful comparisons at this level?
(Yes/No)

Yes.

Data Sources: Are data sources available to support reporting at this level?

No.

Sample Size: What is the typical sample size available for each unit at this level? What proportion of units at this level of aggregation can achieve an acceptable minimum sample size?

This information is not available.

In Use: Have measure results been reported at this level previously?

No.

Reliability & Validity: Is there published evidence about the reliability and validity of the measure when reported at this level of aggregation?

No.

Unintended consequences: What are the potential unintended consequences of reporting at this level of aggregation?

This information is not available.

Other geographic level: Can compare other geographic regions (e.g., MSA, HRR)

Intended use: Is measure intended to support meaningful comparisons at this level?
(Yes/No)

Yes.

Data Sources: Are data sources available to support reporting at this level?

No.

Sample Size: What is the typical sample size available for each unit at this level? What proportion of units at this level of aggregation can achieve an acceptable minimum sample size?

This information is not available.

In Use: Have measure results been reported at this level previously?

No.

Reliability & Validity: Is there published evidence about the reliability and validity of the measure when reported at this level of aggregation?

No.

Unintended consequences: What are the potential unintended consequences of reporting at this level of aggregation?

This information is not available.

Medicaid or CHIP Payment model: Can compare payment models (e.g., managed care, primary care case management, FFS, and other models)

Intended use: Is measure intended to support meaningful comparisons at this level? (Yes/No)

Yes.

Data Sources: Are data sources available to support reporting at this level?

No.

Sample Size: What is the typical sample size available for each unit at this level? What proportion of units at this level of aggregation can achieve an acceptable minimum sample size?

This information is not available.

In Use: Have measure results been reported at this level previously?

No.

Reliability & Validity: Is there published evidence about the reliability and validity of the measure when reported at this level of aggregation?

No.

Unintended consequences: What are the potential unintended consequences of reporting at this level of aggregation?

This information is not available.

Health plan: Can compare quality of care among health plans.

Intended use: Is measure intended to support meaningful comparisons at this level? (Yes/No)

Yes.

Data Sources: Are data sources available to support reporting at this level?

No.

Sample Size: What is the typical sample size available for each unit at this level? What proportion of units at this level of aggregation can achieve an acceptable minimum sample size?

This information is not available.

In Use: Have measure results been reported at this level previously?

No.

Reliability & Validity: Is there published evidence about the reliability and validity of the measure when reported at this level of aggregation?

No.

Unintended consequences: What are the potential unintended consequences of reporting at this level of aggregation?

This information is not available.

Provider Level

Individual practitioner: Can compare individual health care professionals

Intended use: Is measure intended to support meaningful comparisons at this level?

(Yes/No)

Yes.

Data Sources: Are data sources available to support reporting at this level?

Yes.

Sample Size: What is the typical sample size available for each unit at this level? What proportion of units at this level of aggregation can achieve an acceptable minimum sample size?

This information is not available.

In Use: Have measure results been reported at this level previously?

No.

Reliability & Validity: Is there published evidence about the reliability and validity of the measure when reported at this level of aggregation?

No.

Unintended consequences: What are the potential unintended consequences of reporting at this level of aggregation?

This information is not available.

Provider Level

Hospital: Can compare hospitals

Intended use: Is measure intended to support meaningful comparisons at this level?

(Yes/No)

Yes.

Data Sources: Are data sources available to support reporting at this level?

No.

Sample Size: What is the typical sample size available for each unit at this level? What proportion of units at this level of aggregation can achieve an acceptable minimum sample size?

This information is not available.

In Use: Have measure results been reported at this level previously?

No.

Reliability & Validity: Is there published evidence about the reliability and validity of the measure when reported at this level of aggregation?

No.

Unintended consequences: What are the potential unintended consequences of reporting at this level of aggregation?

This information is not available.

Provider Level

Practice, group, or facility: Can compare: (i) practice sites; (ii) medical or other professional groups; or (iii) integrated or other delivery networks**

Intended use: Is measure intended to support meaningful comparisons at this level?

(Yes/No)

Yes.

Data Sources: Are data sources available to support reporting at this level?

Yes.

Sample Size: What is the typical sample size available for each unit at this level? What proportion of units at this level of aggregation can achieve an acceptable minimum sample size?

This information is not available.

In Use: Have measure results been reported at this level previously?

No.

Reliability & Validity: Is there published evidence about the reliability and validity of the measure when reported at this level of aggregation?

No.

Unintended consequences: What are the potential unintended consequences of reporting at this level of aggregation?

This information is not available.

Section 10. Understandability

CHIPRA states that the core set should allow purchasers, families, and health care providers to understand the quality of care for children. Please describe the usefulness of this measure toward achieving this goal. Describe efforts to assess the understandability of this measure (e.g., focus group testing with stakeholders).

The PCPI has worked collaboratively on this measure set with the PCPI-Consumer Purchaser (CPP) Panel. The panel strongly supports this measure and applauds the inclusion of it at the level of the individual clinician. In addition, the workgroup comprised member representatives from consumer groups, patient advocacy groups, and a health plan.

Section 11. Health Information Technology

Please respond to the following questions in terms of any health information technology (health IT) that has been or could be incorporated into the measure calculation.

11.A. Health IT Enhancement

Please describe how health IT may enhance the use of this measure.

The use of health IT in the collection and calculation of this measure allows for the clinical data to be used to assess measure results. The use of clinical data is more desirable compared to administrative data due to the increased granularity of information that can be collected.

11.B. Health IT Testing

Has the measure been tested as part of an electronic health record (EHR) or other health IT system?

Yes.

If so, in what health IT system was it tested and what were the results of testing?

A second phase of reliability testing on the measure also is ongoing at the same sites where feasibility testing was conducted. This approach utilizes parallel forms reliability where measure data elements and performance from an automated report from the EHR are compared to those data from a manual review of the EHR—that is, comparison to the gold standard. (See Measure Testing Protocol for PCPI Performance Measures, ama-assn.org/resources/doc/cqi/pcpi-testing-protocol.pdf.)

11.C. Health IT Workflow

Please describe how the information needed to calculate the measure may be captured as part of routine clinical or administrative workflow.

See Section 8.A/Issues in Implementation for workflow discussion.

11.D. Health IT Standards

Are the data elements in this measure supported explicitly by the Office of the National Coordinator for Health IT Standards and Certification criteria (see healthit.hhs.gov/portal/server.pt/community/healthit_hhs_gov__standards_ifr/1195)?

Yes.

If yes, please describe.

We use the following standards in the development of our EHR specifications: The Quality Data Model (QDM), developed by the National Quality Forum, the vocabulary recommendations named by the Health IT Standards Committee (of the Office of the National Coordinator for Health IT)—e.g., SNOMED, RxNorm, LOINC, and also referenced in the CMS Blueprint. The vocabulary standards were used in the specifications that are a part of Stage II of the CMS EHR incentive program (Meaningful Use). Another available standard is the HL7 Health Quality Measure Format (HQMF), an XML-based structured document to express a quality measure specification. The HQMF is used for specifications included in the Meaningful Use program and also references the QDM. The specifications provided with this submission form have not been incorporated into the HQMF eMeasure format; however, the information included in the specifications serves as the foundation for the HQMF—that is, the PCPI electronic specification outlines the requirements to develop the HQMF.

11.E. Health IT Calculation

Please assess the likelihood that missing or ambiguous information will lead to calculation errors.

It is highly likely that missing data or ambiguous information stored in the EHR will lead to calculation errors. The specifications provided for this measure are designed to query the EHR in order to obtain the data required for the measure calculation.

11.F. Health IT Other Functions

If the measure is implemented in an EHR or other health IT system, how might implementation of other health IT functions (e.g., computerized decision support systems in an EHR) enhance performance characteristics on the measure?

These IT functions could make measure recording in the EHR more feasible and reliable, as well as improve performance on the measure and patient outcomes. For example, computerized

decision support with menu drop downs or reminders could be programmed to give providers prompts to provide patients the appropriate services

Section 12. Limitations of the Measure

Describe any limitations of the measure related to the attributes included in this CPCF (i.e., availability of measure specifications, importance of the measure, evidence for the focus of the measure, scientific soundness of the measure, identification of disparities, feasibility, levels of aggregation, understandability, health information technology).

The measure may have limited utilization due to the limited adoption of EHRs, particularly among practices treating the Medicaid population. However, the vocabulary standards used in the specifications are as proposed for Stage II of the CMS EHR incentive program (Meaningful Use), so its usability is expected to be enhanced by increased participation in this program. As adoption of EHRs increases, utilization of this measure should also increase.

Section 13. Summary Statement

Provide a summary rationale for why the measure should be selected for use, taking into account a balance among desirable attributes and limitations of the measure. Highlight specific advantages that this measure has over alternative measures on the same topic that were considered by the measure developer or specific advantages that this measure has over existing measures. If there is any information about this measure that is important for the review process but has not been addressed above, include it here.

This measure should be selected because it expands the core set of measures beyond their current use. The measure will provide a mechanism to help track unnecessary use of episiotomy to prevent adverse maternal outcomes. This measure is of particular importance for CHIPRA in that it is high impact with Medicaid patients and addresses concerns related to both mother and baby. Additionally, since this measure has full eSpecifications, it can be a candidate for future inclusion in the EHR Incentive Program for Meaningful Use.

Our EHR specifications follow the standards in the Quality Data Model (QDM), developed by the National Quality Forum, the vocabulary recommendations named by the Health IT Standards Committee (of the Office of the National Coordinator for Health IT)—e.g., SNOMED, RxNorm, LOINC, and also referenced in the CMS Blueprint. The vocabulary standards used in the specifications are a part of Stage II of the CMS EHR incentive program (Meaningful Use).

Section 14: Identifying Information for the Measure Submitter

First Name: Ramesh

Last Name: Sachdeva MD, PhD, MBA, FAAP

Title: Professor of Pediatrics (Critical Care)
Organization: Medical College of Wisconsin
Mailing Address: 9000 W. Wisconsin Avenue, MS-681
City: Milwaukee
State: WI
Postal Code: 53226
Telephone: (414) 266-3022
Email: rsachdeva@chw.org

The CHIPRA Pediatric Quality Measures Program (PQMP) Candidate Measure Submission Form (CPCF) was approved by the Office of Management and Budget (OMB) in accordance with the Paperwork Reduction Act.

The OMB Control Number is 0935-0205 and the Expiration Date is December 31, 2015.

Public Disclosure Requirements

Each submission must include a written statement agreeing that, should U.S. Department of Health and Human Services accept the measure for the 2014 and/or 2015 Improved Core Measure Sets, full measure specifications for the accepted measure will be subject to public disclosure (e.g., on the Agency for Healthcare Research and Quality [AHRQ] and/or Centers for Medicare & Medicaid Services [CMS] websites), except that potential measure users will not be permitted to use the measure for commercial use. In addition, AHRQ expects that measures and full measure specifications will be made reasonably available to all interested parties. "Full measure specifications" is defined as all information that any potential measure implementer will need to use and analyze the measure, including use and analysis within an electronic health record or other health information technology. As used herein, "commercial use" refers to any sale, license or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed or distributed for commercial gain, even if there is no actual charge for inclusion of the measure. This statement must be signed by an individual authorized to act for any holder of copyright on each submitted measure or instrument. The authority of the signatory to provide such authorization should be described in the letter.

AHRQ Pub. No. 14(16)-P009-11-EF
March 2016