Behavioral Health Risk Assessment

Section 1. Basic Measure Information

1.A. Measure Name

Behavioral Health Risk Assessment

1.B. Measure Number

0085

1.C. Measure Description

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

Percentage of patients, regardless of age, who gave birth during a 12-month period seen at least once for prenatal care who received a behavioral health screening risk assessment that includes the following screenings at the first prenatal visit: screening for depression, alcohol use, tobacco use, drug use, and intimate partner violence.

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: 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 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)

Not applicable.

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:

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.**

   Not applicable.

1.G. Numerator Statement

Patients who received the following behavioral health screening risk assessments at the first prenatal visit:

Depression screening

Patients who were screened for depression at the first visit: Questions may be asked either directly by a health care provider or in the form of self-completed paper- or computer-administered questionnaires. Results should be documented in the medical record. Depression screening may include a self-reported validated depression screening tool (e.g., Patient Health

Questionnaire-2 [PHQ-2], Beck Depression Inventory, Beck Depression Inventory for Primary Care, Edinburgh Postnatal Depression Scale [EPDS]).

Alcohol use screening

Patients who were screened for any alcohol use at the first visit.

Tobacco use screening

Patients who were screened for tobacco use at the first visit.

Drug use (illicit and prescription, over the counter) screening

Patients who were screened for any drug use at the first visit.

Intimate partner violence screening

Patients who were screened for intimate partner violence/abuse at the first visit: Questions may be asked either directly by a health care provider or in the form of self-completed paper- or computer-administered questionnaires. Results should be documented in the medical record. Intimate partner violence screening may include a self-reported validated depression screening tool (e.g., Hurt, Insult, Threaten, and Scream [HITS], Woman Abuse Screening Tool [WAST], Partner Violence Screen [PVS], Abuse Assessment Screen [AAS]).

To satisfactorily meet the numerator – ALL screening components must be performed.

1.H. Numerator Exclusions

None.

1.I. Denominator Statement

All patients, regardless of age, who gave birth during a 12-month period seen at least once for prenatal care.

1.J. Denominator Exclusions

None.

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 related documents for full specifications and coding spreadsheets. Below is an overview of our technical specifications process.

The PMCoE Cener of Excellence adopted the PCPI specification process, which places emphasis on developing comprehensive measure specifications for electronic health records (EHRs) and provides 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 within EHRs, as they hold promise for supplying relevant clinical data for measures and providing feedback to physicians and other health care providers that is timely and actionable.

The type of specifications provided 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 attached with this submission form 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; and (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 related documents for the 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 Pediatric Measurement Center of Excellence (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.

Currently, there is a quality gap among pregnant women receiving appropriate screenings for depression, drug and alcohol use, smoking, and violence at prenatal visits. Without appropriate screening, it is difficult to assess the number of women who are risk during pregnancy and are putting their babies at risk. This is an important area of focus for measurement and provides a significant area for quality improvement that has implications for mothers, babies, and providers, such as pediatricians.

Clinical depression is common among reproductive-age women and is the leading cause of disability in U.S. women each year. Between 14 and 23 percent of pregnant women will experience depression symptoms during pregnancy, and an estimated 5 to 25 percent of women will have postpartum depression. Studies have shown that untreated maternal depression negatively affects an infant's cognitive, neurologic, and motor skill development. A mother's untreated depression can also negatively impact older children's mental health and behavior. During pregnancy, depression can lead to preeclampsia, preterm delivery, and low birth weight (ACOG, 2010).

In 2002, the U.S. Preventive Services Task Force reviewed evidence about the accuracy of screening instruments in identifying depressed adults. There is little evidence to recommend one screening method over another; therefore, clinicians may choose the method most consistent with their personal preference, the patient population being served, and the practice setting (USPSTF, 2009). Alcohol and substance abuse in pregnant women have been linked to a variety of adverse outcomes for both the mother and her newborn. Besides birth-related, short-term adverse effects, substance use during pregnancy also can lead to long-term developmental problems in the child. Screening pregnant women for alcohol use has become increasingly important because new research indicates that even low levels of prenatal alcohol exposure can negatively affect the developing fetus. Adverse effects of prenatal alcohol exposure can range from subtle developmental problems, or fetal alcohol effects, to full-blown fetal alcohol syndrome. In addition, certain neurobehavioral outcomes associated with prenatal alcohol exposure can persist in the affected person into adolescence (Sampson, Bookstein, Barr, et al. 1994) and adulthood (Kelly, Day, Streissguth, 2000). According to new studies, even low levels of prenatal alcohol exposure can negatively affect the developing fetus, thereby increasing the importance of identifying women who drink during pregnancy. In response, researchers have developed several simple alcohol screening instruments for use with pregnant women. These instruments, which can be administered quickly and easily, have been evaluated and found to be effective. Because of the potential adverse consequences of prenatal alcohol exposure, short screening questionnaires are

worthwhile preventive measures when combined with appropriate followup. Women abused during pregnancy are more likely to be depressed, suicidal, and experience pregnancy complications and poor outcomes, including maternal and fetal death.

American College of Obstetricians of and Gynecologists (ACOG). Committee Opinion No. 453. Screening for depression during and after pregnancy. Obstest Gynecol 2010; 115: 394-5.

Kelly SJ, Day N, Streissguth AP. Effects of prenatal alcohol exposure on social behavior in humans and other species. Neurotoxicol Teratol 2000; 22(2):143-9.

Sampson PD, Bookstein FL, Barr HM, et al. Prenatal alcohol exposure, birthweight, and measures of child size from birth to age 14 years. Am J Public Health 1994; 84(9):1421-8.

U.S. Preventive Services Task Force. Screening for Depression. Rockville, MD: Agency for Healthcare Research and Quality; 2009. Available at http://www.ahrq.gov/professionals/prevention-chronic-care/healthier-pregnancy/preventive/depression.html. Accessed November 4, 2015.

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 identify patients with drug, alcohol, or smoking problems, as well as depression and abuse, which may help prevent adverse neonatal outcomes. Women abused during pregnancy are more likely to be depressed, suicidal, and experience pregnancy complications and poor outcomes, including maternal and fetal death. 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).

Currently, there are no measures that assess pregnant women for depression, alcohol and drug use, tobacco use, and intimate partner violence. We believe that there is a quality gap in screening and followup for women at risk, particularly those in the Medicaid population.

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 :** Yes.
b. **Care Setting – inpatient :** No.
c. **Care Setting – other – please specify :** No.
d. **Service – preventive health, including services to promote healthy birth :** Yes.
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 :** No.
i. **Measure Topic – clinical quality :** Yes.
j. **Measure Topic – patient safety :** Yes.
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) :

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.

Evidence Behind the measure: The evidence behind smoking, drug use, and alcohol use during pregnancy and its link to increased risk of adverse outcomes for mothers and babies includes clinical practice guidelines and numerous published research studies.

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 and Prevention.
- United States Preventive Services Task Force.
- Veterans Administration/Department of Defense Clinical Practice Guideline for Pregnancy Management.
- Society of Obstetricians and Gynecologists of Canada.

Numerous research studies have assessed the lack of screening for depression and alcohol and substance use among pregnant women. A 2003 report demonstrated the prevalence of depressive symptomatology during pregnancy when seen in obstetric settings, the extent of treatment in this population, and specific risk factors associated with mood symptoms in pregnancy. A total of 3,472 pregnant women age 18 and older were screened while waiting for their prenatal care visits in 10 obstetric clinics using a brief (10 minute) screening questionnaire. This screen measured demographics, tobacco and alcohol (TWEAK problem alcohol use screening measure), and depression measures, including the Center for Epidemiological Studies-Depression scale (CES-D), use of antidepressant medications, past history of depression, and current treatment (i.e., medications, psychotherapy, or counseling) for depression. Of those women screened, 20 percent (n = 689) scored above the cutoff score on the CES-D, and only 13.8 percent of those women reported receiving any formal treatment for depression. Past history of depression, poorer overall health, greater alcohol use consequences, smoking, being unmarried, unemployment, and lower educational attainment were significantly associated with symptoms of depression during pregnancy. These data show that a substantial number of pregnant women screened in obstetric settings have significant symptoms of depression, and most of them are not being monitored in treatment. As elevations in depressive symptomatology have been associated with adverse maternal and infant outcomes, further study of the impact of psychiatric treatment in gravid women is essential.

The U.S. Preventive Services Task Force (USPSTF) reviewed evidence about the accuracy of screening instruments in identifying depressed adults in 2002. Many formal screening tools are available, including instruments designed specifically for older adults. There is little evidence to recommend one screening method over another; therefore, clinicians may choose the method most consistent with their personal preference, the patient population being served, and the practice setting. (USPSTF, 2009)

By integrating routine screening and treatment for substance use, including alcohol and cigarette smoking, into the prenatal care system, the health outcomes of mothers and their babies can be significantly improved, according to a retrospective study conducted by a large U.S. health care organization. The study examined the records of nearly 50,000 pregnant women who went through the prenatal substance use screening between 1999 and 2003. They found that women who were screened positive, assessed by the specialist, and treated for substance use had significantly better birth-related outcomes than those who screened positive but turned down assessments and/or treatment by the Early Start specialist. The birth-related benefits were seen in both the mothers and the newborns. The risk of having a preterm delivery, placental abruption, and intrauterine fetal death (stillbirth) were all significantly reduced. The babies born to mothers who underwent the Early Start program had lower risks of requiring neonatal-assisted ventilation and having low birthweight. Of the women included in the study, 2,073 were positive for alcohol, smoking, or substance use at screening and received an assessment and at least one followup appointment with a specialist; 1,203 were screened positive, assessed by the specialist, and declined a followup appointment; and 156 were screened positive but received neither assessment nor followup. The other 46,000 women who had negative results at screening served as the control group.

The workgroup reviewed multiple evidence-based clinical practice guidelines for supporting evidence for this measure. The following guideline statements were used as a basis for this measure.

Depression Screening

- A social and mental health history should be completed on all new prenatal patients.

- Routine depression screening is recommended for all patients in clinical practices that have systems in place to assure effective diagnosis, treatment, and followup.

Depression Screening Weeks 6-8, 28 (Veterans Administration/Department of Defense Clinical Practice Guideline for Pregnancy Management, 2009)

- Women should be screened for depression during their first contact with obstetric health care services, at week 28, and at the postpartum visit.

- Depression screening should be performed using a standardized screening tool, such as the Edinburgh Postnatal Depression Scale (EDPS) or the PHQ-2.

- Women should be asked early in pregnancy if they have had any previous psychiatric illnesses. If they have a past history of serious psychiatric disorder, they should be referred for a psychiatric assessment during the antenatal period (USPSTF, 2009).

- All positive screening tests should trigger full diagnostic interviews that use standard diagnostic criteria to determine the presence or absence of specific depressive disorders, such as major depressive disorder (MDD) or dysthymia.

- The severity of depression and co-morbid psychological problems (for example, anxiety, panic attacks, or substance abuse) should be addressed.

Alcohol and Drug Use Screening

The USPSTF strongly recommends (B Recommendation) screening and behavioral counseling interventions to reduce alcohol misuse by adults, including pregnant women, in primary care settings (USPSTF, 2004).

The USPSTF strongly recommends (A Recommendation) that clinicians screen all adults for tobacco use and provide tobacco cessation interventions for those who use tobacco products. (USPSTF, 2003)

Intimate Partner Violence Screening

The Society of Obstetricians and Gynaecologists of Canada recommends (SOCG, 2005):

1. Providers should include queries about violence in the behavioral health assessment of new patients, at annual preventive visits, as a part of prenatal care, and in response to symptoms or conditions associated with abuse (B).

B: There is fair evidence to support the recommendation for use of a diagnostic test, treatment, or intervention.

Summary statement

1. At least three systematic reviews of "screening" for intimate partner violence (IPV) have found insufficient evidence to recommend for or against routine screening. Asking women about violence is not a screening intervention: victims are not asymptomatic; disclosure is not a test result, it is a voluntary act, and the presence or absence of violence is not under the victim's control; and most interventions required to protect and support survivors are societal, not medical.(I).

I: Evidence obtained from at least one properly designed randomized controlled trial.

American College of Obstetricians and Gynecologists, 2012.

Obstetrician–gynecologists are in the unique position to provide assistance for women who experience IPV because of the nature of the patient–physician relationship and the many opportunities for intervention that occur during the course of annual examinations, family planning, pregnancy, and followup visits for ongoing care. (Not rated)

Screening all patients at various times is also important because some women do not disclose abuse the first time they are asked. Health care providers should screen all women for IPV at periodic intervals, such as annual examinations and new patient visits. Signs of depression, substance abuse, mental health problems, requests for repeat pregnancy tests when the patient does not wish to be pregnant, new or recurrent sexually transmitted infections (STIs), asking to be tested for an STI, or expressing fear when negotiating condom use with a partner should prompt an assessment for IPV. (Not rated)

Screening for IPV during obstetric care should occur at the first prenatal visit, at least once per trimester, and at the postpartum checkup. (Not rated)

Goler NC , Armstrong MA, Taillac CJ, et al. Substance abuse treatment linked with prenatal visits improves perinatal outcomes: a new standard. J Perinatol 2008; 28:597-603. doi:10.1038/jp.2008.70.

Institute for Clinical Systems Improvement (ICSI). Major Depression in Adults in Primary Care. Bloomington, MN: Institute for Clinical Systems Improvement (ICSI); May 2008.

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 an 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 and other providers (Adams, Mehrotra, McGlynn, 2010; Physician Cost Profiling, 2010; Scholle, Roski, Adams, et al, 2008). The signal is the variability in measured performance that can be explained by real differences in physician performance, and the No.ise 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 (Nunnelly, Bernstein, 1994).
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: (1) at a minimum number of quality reporting events per physician and (2) 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 National Quality Forum (NQF) endorsement (NQF, 2012).

The SNR reliability testing for this measure is underway. We are currently producing the automated report from the EHR and will be completing reliability testing analysis when those data become available. We expect to have reliability testing and performance results prior to the SNAC meeting in September. The analysis will provide results on measure reliability, overall measure performance, the distribution of performance rates, and performance stratified by patient race, ethnicity, preferred language, socioeconomic status, and demographic variables. The structure of the results is the same as that included in our submission of the PMCoE/AMA-PCPI c-section and episiotomy measures.

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.)

Adams JL, Mehrotra A, McGlynn EA. Estimating Reliability and Misclassification in Physician Profiling, Santa Monica, CA: RAND Corporation, 2010. Available at http://www.rand.org/pubs/technical_reports/TR863. Accessed November 4, 2015.

NQF Removes Time-Limited Endorsement for 13 Measures; Measures Now Have Endorsed Status. Washington, DC: National Quality Forum; 2012. Available at http://www.qualityforum.org/News_And_Resources/Press_Releases/2012/NQF_Removes_Time-Limited_Endorsement_for_13_Measures;_Measures_Now_Have_Endorsed_Status.aspx. Accessed November 4, 2015.

Nunnelly J, Bernstein I. Psychometric Theory. 3rd ed. New York, NY: McGraw-Hill; 1994.

Physician cost profiling--reliability and risk of misclassification. N Engl J Med. 2010 Mar 18;362(11):1014-21. http://www.nejm.org/doi/pdf/10.1056/NEJMsa0906323.

Scholle SH, Roski J, Adams JL, et al. Benchmarking physician performance: reliability of individual and composite measures. Am J Manag Care. 2008, 14:833-838. Available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2667340/pdf/nihms-99203.pdf.

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 subject matter expert panel consisted of 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 also by 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. Other external review groups (e.g., focus groups) may be convened if there are any remaining concerns related to the content validity of the measures.

The expert panel members also assessed the measure 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 Behavioral Health Risk Assessment measure, the mean score was 4.46; 84.6 percent (11/13) of respondents agree or strongly agree that the scores obtained from the measure as specified will accurately differentiate quality across providers; and 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 a Supplemental Data Element 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, consistent with recent national efforts to standardize the collection of race and ethnicity data. We have included these variables as recommended data elements to be collected in the measure specifications.

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.

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 Tool

The PMCoE Cener of Excellence adopted the AMA-PCPI testing methodology which uses the Data Element Table (DET)© Tool[a] 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 it is feasible for each site to collect the data for the measures. It is structured to collect metadata 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.

[a] Data Element Table Tool: copyright 2013, American Medical Association. All rights reserved. This tool and the information contained therein may not be reproduced or distributed and may only be used for collecting data in connection with an agreement with the American Medical Association. This tool is provided "as is" without warranty of any kind.

- *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 specific site. For some sites, all of the measures in the Perinatal/Prenatal Measurement Set may be collected, for others it may be only a few.

Once the completed DET was submitted by the test site, 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 an urban, tertiary care hospital.

All nine of the data elements can be captured in code or text format.

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 which 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."

- "Non-feasible. 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."

The test site rated the technical and implementation feasibility of the measure as "Non-feasible. Unable to do today." The site reported that the data for this measure is not being captured in their inpatient HER, and that they do not know whether the data is captured reliably in the outpatient record. Unavailability of the data in an inpatient EHR would not affect feasibility, however, since the measure is specified for ambulatory care settings. The site indicated that making the measure feasible would require a change in outpatient documentation.

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.

The development of the measure was completed in a short time; earlier this year; hence there was 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 AMA-PCPI has worked collaboratively on this measure set with the AMA-PCPI-Consumer Purchaser Panel (CPP), which comprised representatives from the patient, consumer, and purchaser communities. The panel strongly supports this measure addressing pertinent issues of behavioral health and applauds the inclusion of it at the level of the individual clinician. The CPP states this important measure can help to identify at-risk patients and provide treatment and followup during and after pregnancy. In addition, the work group included 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 of 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 NQF, 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 consistent with those recommendations proposed for 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 serve 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 health 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 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. 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 NQF, 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:

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.

The signed written statement was submitted

AHRQ Pub. No. 14(16)-P009-8-EF
December 2015