

**Measure Testing Protocol for Physician Consortium for Performance  
Improvement<sup>®</sup> (PCPI<sup>™</sup>) Performance Measures**

**First Approved by the PCPI October 5, 2007  
Updated December, 2010**

# Measure Testing Protocol for Physician Consortium for Performance Improvement Performance Measures

## OVERVIEW

### 1. Background

The Physician Consortium for Performance Improvement<sup>®</sup> (PCPI<sup>™</sup>) has long viewed measure testing as a critical step in the measure development process. In 2006, the Measures Implementation and Evaluation (MIE) Advisory Committee of the PCPI was charged with developing a protocol: 1) devising a standardized methodology to test performance measures and 2) guiding measure testing activities. The MIE developed the first *Measure Testing Protocol for PCPI Performance Measures* (hereafter referred to as the *PCPI Measure Testing Protocol*, or the Protocol) and it was approved by the PCPI October 5, 2007. This document, a 2010 update, replaces the original Protocol. The PCPI, along with a number of collaborators, is carrying out testing of the PCPI performance measures to determine the feasibility, validity, and reliability of the physician performance measures based on this Protocol.

### 2. Purpose and Approach

The key purposes of the *PCPI Measure Testing Protocol* are to:

- Assist the PCPI in prioritizing measure testing activities by identifying key scientific areas for testing that would collectively constitute a comprehensive evidence base *for each PCPI measure*,<sup>1</sup> and address the scientific attributes that the PCPI has previously deemed desirable of performance measures.<sup>2</sup>
- Provide research recommendations where appropriate and possible to help ensure an adequate level of quality for research to be included within the evidence base *for each PCPI performance measure*.

To achieve these objectives, the PCPI MIE recommends a prioritized approach to testing the PCPI measures. **Priority I** covers testing activities that can be carried out in the short-run, and includes four testing areas: Needs Assessment (Testing Area 1); Feasibility and Implementation (Testing Area 2); Reliability (Testing Area 3); and Validity (Testing Area 4), including face and construct validity. **Priority II** covers testing activities that are likely to extend over the long run and are of a more applied nature. Priority II testing areas include: the Predictive Validity aspect

---

<sup>1</sup> The evidence base *for a performance measure* is distinct from the evidence base *for a process of care*. The evidence base *for a performance measure* includes: 1) evidence justifying the existence of a given measure (eg, gaps in care, variations in care associated with the process of care around which a performance measure has been developed); 2) evidence regarding the feasibility of a performance measure, or empirical research describing the implementation of a given measure; 3) evidence regarding the reliability of a given measure; 4) evidence regarding the validity of a given measure; and 5) evidence regarding unintended consequences associated with the use of a given measure. The evidence base *for a process of care* concerns clinical studies of efficacy and effectiveness regarding the use of a specific process of care and patient outcomes. This protocol is concerned with building an evidence base *for PCPI performance measures*.

<sup>2</sup> Physician Consortium for Performance Improvement. Desirable Attributes of Performance Measures: A Consensus Document from The American Medical Association, The Joint Commission on Accreditation of Healthcare Organizations, and The National Committee for Quality Assurance. April 19, 1999.

of Validity (Testing Area 4) Unintended Consequences (Testing Area 5); and Applications of PCPI Performance Measures (Testing Area 6).

For a summary of all testing recommendations in this Protocol, readers are referred to Appendix A.

### **3. Relevance to National Quality Forum Endorsement Testing Contingency**

This Protocol outlines what can be considered an “ideal” approach for testing performance measures, and it describes a comprehensive set of testing activities that the PCPI MIE considers for all performance measures **for the purpose of building a sound evidence base that attests to the scientific merit of each PCPI measure.** Measures that have not been formally tested are only eligible to receive a “time-limited” endorsement from NQF. This time-limited endorsement requires measure developers to provide a testing plan and testing results within 1 year to retain NQF endorsement.

### **4. Implementation of the PCPI Measure Testing Protocol**

It is not the expectation of the PCPI that *all* individual tests for each measure identified in this document be carried out for each individual PCPI measure.

#### ***An Ideal Approach to Testing***

The following PCPI Measure Testing Protocol outlines a comprehensive, broad-based approach covering a full spectrum of testing approaches, without implying that all aspects of the Protocol should be applied in full to each testing project. This Protocol document outlines what an ideal, comprehensive testing of each PCPI measure could include, before considering what subset of tests and test conditions may be considered “acceptable” for practical purposes. The ideal conceptualization of comprehensive testing for each PCPI measure includes:

- Tests in all 6 Testing Areas;
- Tests of feasibility/implementation, reliability, validity, and unintended consequences conducted for each data collection modality for which PCPI measures have been developed; and
- Tests of feasibility/implementation, reliability, validity, and unintended consequences conducted in a variety of practice settings including (eg, for ambulatory measures, tests should be conducted in solo practices, medium sized practices, large practices, safety-net ambulatory practices).

#### ***A Practical Approach to Testing: Criteria for Judging a PCPI Measure as ‘Tested’***

The MIE recognizes that an ideal concept of comprehensive testing provides little decision-making guidance for practical purposes. To this end, the MIE recommends that stakeholder evaluation of PCPI measures be based on evidence from a subset of the tests identified in this Protocol. The MIE believes it is the prerogative of each stakeholder to form their own evaluative criteria regarding which type and combination of tests identified in this Protocol must be carried out to best serve each stakeholder’s purpose for using or endorsing the PCPI measures. A list of example testing projects is available in Appendix C.

## **AUTHORS**

This Protocol was developed by the **Measures Implementation and Evaluation Advisory Committee** of the Physician Consortium for Performance Improvement (PCPI). The earlier Protocol was approved by the PCPI at its October 5, 2007 meeting.

We thank the members of that Advisory Committee for their expertise and contribution:

### **2010 Committee**

Robert Bonow, MD (Chair)  
David Baker, MD, MPH, FACP  
Julie Cerese, RN, MSN  
Joseph Drozda, MD  
R. Adams Dudley, MD, MBA  
Fred Edwards, MD, MS  
Christine M. Goertz, D.C., PhD  
Cindy P. Helstad, PhD, RN  
Jeffrey A. Linder, MD, MPH, FACP  
Barbara McNeil, MD, PhD  
Mark Metersky, MD, FCCP  
Martha Radford, MD, FACC, FAHA  
David Shahian, MD, FACS  
Aamir Siddiqui, MD, FACS  
Josie R. Williams, MD, MMM, RN, BS  
Helen Burstin, MD, MPH (NQF liaison)

### **AMA Staff 2010**

Keri Christensen, MS  
Gregory Wozniak, PhD  
Karen Kmetik, PhD

## COMMENTS AND QUESTIONS

For additional information regarding this Protocol, please contact:

Keri Christensen, MS  
American Medical Association  
515 North State Street  
Chicago, Illinois 60654  
[keri.christensen@ama-assn.org](mailto:keri.christensen@ama-assn.org)  
(312)464-4805

Joanne Cuny, RN, MBA  
American Medical Association  
515 North State Street  
Chicago, Illinois 60654  
[Joanne.cuny@ama-assn.org](mailto:Joanne.cuny@ama-assn.org)  
(312) 464-4420

Gregory Wozniak, PhD  
American Medical Association  
515 North State Street  
Chicago, Illinois 60654  
[greg.wozniak@ama-assn.org](mailto:greg.wozniak@ama-assn.org)  
(312) 464-4594

**Measure Testing Protocol for Physician Consortium for Performance Improvement  
Performance Measures**

**CONTENTS**

|  | Page |
|--|------|
| Introduction   | 1    |
| <b><u>PRIORITY I TESTING</u></b>   |      |
| Section I. <a href="#">Testing Area 1: Needs Assessment</a>  | 5    |
| Section II. <a href="#">Testing Area 2: Feasibility and Implementation</a><br>2.3 Pilot Testing                              | 9    |
| Section III. <a href="#">Testing Area 3: Reliability</a>   | 22   |
| Section IV. <a href="#">Testing Area 4: Validity:</a><br>4.2 Face Validity<br>4.3 Content Validity<br>4.4 Construct Validity | 45   |
| <b><u>PRIORITY II TESTING</u></b>  |      |
| Section IV. <a href="#">Testing Area 4: Validity</a><br>continued       4.5 Predictive Validity                              | 49   |
| Section V. <a href="#">Testing Area 5: Unintended Consequences</a>   | 56   |
| Section VI. <a href="#">Applications</a>   | 63   |
| <b><u>APPENDICES</u></b>   |      |
| APPENDIX A <a href="#">List of Key Recommendations for Measure Testing</a>   | 67   |
| APPENDIX B <a href="#">PCPI Performance Measures as of 2010</a>  | 69   |
| APPENDIX C <a href="#">Example Testing Efforts for PCPI Measures</a>   | 70   |

## INTRODUCTION

### *Introduction*

The necessity of empirically establishing the feasibility, validity, and reliability of the PCPI performance measures has long been acknowledged. In 2006, the PCPI convened the Measures Implementation and Evaluation (MIE) Advisory Committee and charged the committee with the development of a protocol for testing the PCPI performance measures. This Protocol was to provide recommendations to guide the formation of a thorough evidence base for each PCPI measure. This document, approved by the PCPI on October 5, 2007, and updated September, 2010, contains the set of recommendations developed by the PCPI MIE to guide the formation of a solid testing evidence base for each PCPI physician measure that has already been developed. The PCPI MIE recommendations for PCPI measures that are under development are included in the sections regarding Testing Area 2: Feasibility and Implementation, and Testing Area 3: Reliability.

The key purposes of the *PCPI Measure Testing Protocol* are to:

1. Assist the PCPI in prioritizing measure testing activities by identifying key scientific areas for testing that would collectively constitute a comprehensive evidence base for each PCPI measure, and that address the scientific attributes that the PCPI has previously deemed desirable of performance measures.<sup>3</sup>
2. Provide research recommendations, where appropriate and where possible, to help ensure an adequate level of quality for research to be included within the evidence base for each PCPI performance measure.

### *Users and Uses of This Protocol*

The intended Users of this Protocol are members of the PCPI who will be reviewing the empirical testing evidence base for each PCPI performance measure. However, it is also anticipated that national measure-endorsing organizations, researchers, and users of PCPI measures may consult this document. Measures are designed for use by any physician or other health care professional, where appropriate, who manages the care of a patient for a specific clinical condition, undergoing a particular diagnostic or therapeutic procedure or for prevention of that condition.

PCPI measures may be used by:

*PCPI Members.* It is the intent of this Protocol to provide a basis upon which PCPI members will be able to assess the completeness of the evidence base for each performance measure, as well as criteria against which to evaluate the quality of research considered as part of the evidence base for a performance measure. Pending review of the evidence base for a performance measure, the PCPI will consider the preponderance of evidence for, or against each measure with respect to each of the 6 Testing Areas described below.

---

3 Physician Consortium for Performance Improvement. Desirable Attributes of Performance Measures: A Consensus Document from The American Medical Association, The Joint Commission on Accreditation of Healthcare Organizations, and The National Committee for Quality Assurance. April 19, 1999.

*Measure-Endorsing Organizations.* Measure-endorsing organizations may find this Protocol a useful resource against which to consider the completeness and quality of PCPI performance measures. By making the PCPI testing approach publicly transparent, this Protocol will serve to facilitate a dialogue between PCPI and other measure developers and/or measure-accrediting organizations who may have their own definitions and standards for testing measures.

*Researchers.* Researchers interested in studying quality measurement and improvement using PCPI or other performance measures may find this Protocol useful in identifying areas for investigation – particularly if researchers intend to collaborate with the PCPI, and/or if researchers wish their study to be considered for inclusion in the evidence base for a performance measure. This Protocol may also inform the design of studies within the 6 Testing Areas described below.

*Users of PCPI Measures.* Users or potential users considering adoption of PCPI measures in performance reporting programs (for example, payers, including the Centers for Medicare and Medicaid Services) may find this Protocol a helpful resource in determining that a measure is not only feasible, but scientifically sound, and has predictive power and practical application(s).

Regardless of User, this Protocol may be applied retrospectively to existing research on a current PCPI performance measure to determine if the research: (a) falls within one of the 6 Testing Areas described below; (b) satisfies the scope conditions; and (c) meets the relevant research recommendations/standards delineated in Sections I-VI of this Protocol. This Protocol may be applied prospectively to guide the selection of an area for investigation (the proposed study falls within one of the 6 Testing Areas and satisfies any additional scope conditions within the Testing Area), as well as to inform study design (adheres to the research recommendations for the Testing Area).

### ***Scope of This Protocol***

The scope of this Protocol encompasses empirical research for consideration as part of the evidence base for PCPI performance measures that satisfy the following conditions:

- The measure has already been developed by the PCPI, unless otherwise noted and discussed in this Protocol.
- The measure is a process-of-care performance measure. To date, the majority of PCPI measures are process measures. As additional types of measures (eg, composite, outcome, and/or structural measures) are developed, we will revise this Protocol if necessary to address the testing of those measures appropriately.
- The measure refers to ‘underuse’ – ie, a service or pattern of recommended care that has been consistently demonstrated in clinical/healthcare research to be underprovided or underutilized. Testing guidelines in this Protocol may also apply to the evaluation of appropriate use measures.

### ***A Complete Evidence Base for an Existing PCPI Performance Measure.***

The PCPI MIE recommends that all performance measures developed under the auspices of the Physician Consortium for Performance Improvement® (PCPI) should be tested to:

1. Assess the justification for existence of a performance measure (Testing Area 1)
  - Why are we interested in measuring this aspect of care?
2. Assess the feasibility of a performance measure (Testing Area 2)
  - Can we measure this aspect of care with reasonable cost and level of effort?
3. Evaluate and document the reliability performance measure (Testing Area 3)
  - Can we measure this aspect of care well?
4. Evaluate and document the validity of a performance measure (Testing Area 4)
  - Does high performance lead to better patient outcomes?
5. Investigate potential unintended consequences of a performance measure (Testing Area 5)
  - What are the potential adverse effects of measuring this aspect of care?
6. Applications (Testing Area 6)
  - What are the potential benefits of measuring this aspect of care?

A testing evidence base for a specific PCPI performance measure may be deemed *complete* if empirical evidence covers all six testing areas, and if the evidence is deemed by the PCPI to be of sufficient quality according to the PCPI MIE research recommendations in this Testing Protocol.

***Recommendation.* All PCPI performance measures should be tested in each of the following testing areas:**

- Needs Assessment (Testing Area 1)
- Feasibility (Testing Area 2)
- Reliability (Testing Area 3)
- Validity (Testing Area 4)
- Unintended Consequences (Testing Area 5)
- Applications (Testing Area 6)

### ***Organization of the Protocol***

The remainder of this Protocol discusses each Testing Area in detail, and provides research recommendations to guide the selection and development of a solid testing evidence base for each PCPI performance measure.

The PCPI MIE recommends a prioritized approach to testing the PCPI measures, which is reflected in the organization of this Protocol. **Priority I** covers testing activities that can be carried out in the short-run, and includes four testing areas: Needs Assessment (Testing Area 1); Feasibility and Implementation (Testing Area 2); Reliability (Testing Area 3); and Validity (Testing Area 4). **Priority II** covers testing activities that are likely to extend over the long run and are of a more applied nature. Priority II testing areas include: the Predictive Validity aspect of Validity (Testing Area 4), Unintended Consequences (Testing Area 5); and Applications of PCPI Performance Measures (Testing Area 6).

For illustration and clarification, this Protocol will draw upon existing PCPI performance process measures as examples throughout this document. The American College of Cardiology/American Hospital Association/PCPI Beta-Blocker Performance Measure from the Coronary Artery Disease measure set is defined as the percentage of patients with prior myocardial infarction (MI) who were prescribed beta-blocker therapy. The components of this measure (i.e., numerator, denominator, and exception criteria) are defined in Box 0-1.

#### **Box 0-1. ACC/AHA/PCPI Beta-Blocker Performance Measure**

*(from the Coronary Artery Disease measurement set)*

##### **Percentage of patients with prior MI at any time who were prescribed beta-blocker therapy**

***Numerator:*** Patients who were prescribed beta blocker therapy

***Denominator:*** All patients with CAD who also have prior MI at any time > 18 years of age

***Denominator inclusion:*** Patients with CAD and prior MI

***Denominator exception:***

Documentation of medical reason(s) for not prescribing beta blocker therapy;

Documentation of patient reason(s) for not prescribing beta blocker therapy

## SECTION I. TESTING AREA 1: NEEDS ASSESSMENT

### Section I Outline

- 1.1 [General Introduction](#)
- 1.2 [Research Recommendation for Demonstrating a Need for Existing Measures](#)
  - 1.2.1. [Acceptable Types of Evidence](#)
- 1.3 [Minimal Standards for Documenting a Clinically Significant Gap in Care](#)

## ***1.1. General Introduction***

The PCPI endeavors to develop measures in clinical domains for which a gap or variation in care exists. In the context of this discussion, a *gap in care* (or *variation in care*) refers to observed deviation (or observed patterns of deviation) in care from established norms or standards of care as formalized in clinical practice guidelines. Gaps in care may be manifested by underuse, overuse, or misuse of health service or treatments. Following the Institute of Medicine National Roundtable on Health Care Quality, we define *underuse* of health services as “the failure to provide a health care service when it would have produced a favorable outcome for a patient.”<sup>4</sup> We define *overuse* of health services as the provision of health services “under circumstances in which its potential for harm exceeds the possible benefit,”<sup>1</sup> or circumstances in which there is zero benefit, regardless of harm. A gap in care is *clinically important* if the deviation from clinical standards/guidelines is strongly implicated in producing avoidable, negative health outcomes.

## ***1.2. Research Recommendation for Demonstrating a Need for Existing Measures***

For an existing measurement set developed by the PCPI, it is recommended that evidence be periodically provided demonstrating:

- The continued existence of a clinically important gap in care and
- The prevalence of gaps in care (underuse or overuse) in the relevant patient population with regard to a specific measure set.

The decision of when to “retire” or remove an individual measure from a measurement set based on empirical evidence that the original gap in care has been significantly decreased is to be made by the PCPI Measure Development Work Group of experts who developed the measure.

***Recommendation 1-1. For existing performance measures, it is recommended that the existence of a gap in care be re-evaluated by the measure developing body or other evaluators at the time that the measurement set is being formally updated (for the PCPI, measures are updated every 3 years).***

The evidence provided as justification for an existing PCPI measure set must include studies of an association between deviation in care and avoidable, poor outcomes. The evidence base that measure developers, such as the PCPI, use to support the efficacy of a given process of care to result in a desired outcome is derived from clinical guidelines. Evidence must also be provided regarding the prevalence of deviations or gaps in care. Evaluators of this evidence (eg, a Work Group of the PCPI or the NQF) will judge whether the association between gaps in care and poor outcomes is sufficiently strong to warrant a performance measure or measure set. Evaluators of this evidence will also judge whether the prevalence of gaps in care is of a sufficient magnitude to warrant a performance or measure set.

---

<sup>4</sup> Chassin MR, Galvin RW, and the National Roundtable on Health Care Quality. The urgent need to improve health care quality. *JAMA*. 1998;280:1000-1005.

### 1.2.1. Acceptable Types of Evidence

Previously published studies demonstrating a gap in care and/or the clinical significance of a gap in care may be cited as evidence supporting the need for an existing PCPI performance measure. If published studies are non-existent or of inadequate quality, studies involving the analysis of appropriate data (such as reviewing data from large health plans to identify a potential quality failure) should be undertaken to quantify the association between a gap in care and avoidable, poor outcomes, and/or to quantify the prevalence of a gap in care.

**Recommendation 1-2. For an existing PCPI measure, acceptable types of evidence for documenting significant gaps in care include: 1) previously published studies in peer-reviewed publications; and 2) secondary analysis of existing data.**

Previously published studies cited as evidence of clinically significant gaps in care should have been published in a peer reviewed journal within the 5 years immediately preceding a measure set's consideration. The data on which the study is based (or the most recent wave of data included in the study) should not predate the time a measure set is being considered by more than 10 years.

### 1.3. Minimal Standards for Documenting a Clinically Significant Gap in Care

The following are standards describing a minimally acceptable body of evidence for a gap in care that may be brought forth to an evaluating body.

**Standard #1: Evidence should be provided concerning the prevalence of a gap in care at the national level.** Studies cited as evidence must be based on data that are representative of the relevant patient population at the national level. If there are no studies based on nationally representative data sets, a group of studies using regional data or other subgroups may be considered if the collection of studies taken together address all segments of the relevant patient population. An example of evidence of a gap in care is shown in Box 1-1.

#### Box 1-1. ACC/AHA/PCPI Beta-Blocker Performance Measure<sup>5</sup>

Evidence of a gap in care for beta-blocker therapy among patients with CAD:

- The 'initial prescribing rate' at discharge was found to be 55% for beta-blockers.
- "Continuity of prescribing" for 5 years was found to be 20% for beta-blockers

<sup>5</sup>Rabus SA, Izzettin FV, Sancur M, Karakaya O, Kargin R, Yakut C. Five-year follow-up of drug utilization for secondary prevention in coronary artery disease. *Pharmacology World and Science*. 2008;30(6)753-758.

**Standard #2. Subgroup analyses should be conducted or cited to ascertain whether variations in the prevalence of a gap in care exist across important clinical or sociodemographic groups of patients.** Subgroup analyses may illuminate whether aggregate rates of underuse or overuse are driven by patterns of care in isolated subgroups. Although PCPI measures may serve as useful tools in efforts to reduce disparities in care, the measures are designed to promote safe and quality care for all relevant patients.

Desirable attributes of studies include the following:

- Adequate statistical power; and
- Data from random samples of providers and patients (where applicable).

## SECTION II. TESTING AREA 2: FEASIBILITY AND IMPLEMENTATION

### Section II Outline

- 2.1. [Feasibility, Implementation and PCPI Performance Measures](#)
  - 2.1.1. [Definitions of Key Terms in this Section](#)
  - 2.1.2. [Recommendation](#)
  - 2.1.3. [Rationale](#)
  - 2.1.4. [Methodology – Pilot testing for Implementation](#)
    - [Figure 1: Current Measure Development Process](#)
    - 2.1.4.1. [Questionnaire Based Pilot Testing](#)
    - 2.1.4.2. [Focus Group Based Pilot Testing](#)
      - [Figure 2: PCPI Measure Development and Testing Cycle](#)
  - 2.1.5. [Methodology – Description of Implementation Strategy](#)
  - 2.1.6. [Methodology – Feasibility of Data Collection](#)
  - 2.1.7. [Methodology – Barriers Analysis](#)
  - 2.1.8. [Methodology – Resource Utilization/Cost Analysis](#)
- 2.2. [Scope of Feasibility and Implementation Testing](#)
  - 2.2.1. [Recommendation](#)
  - 2.2.2. [Rationale](#)

## 2.1. Feasibility, Implementation and PCPI Performance Measures

### 2.1.1. Definitions

The *feasibility* of a PCPI performance measure/measure set refers to the extent to which clinical practices are able to interpret measure definitions and technical specifications, and 1) integrate them into existing workflows and health information systems to collect, manage, and manipulate data elements; 2) compute performance measures; and 3) generate performance reports within a reasonable time frame and budget.

*Implementation* refers to the changes in practice organization that are necessary to create, use, and maintain the capacity to report on a PCPI performance measure or measure set. Examples of organizational changes associated with PCPI performance measure implementation include:

- Acquisition and/or modification of technology
- Changes in physical capital
- Changes in staffing (including role relationships, definitions, and responsibilities)
- Changes in workflow structure and processes
- Changes in practice culture and policies
- Changes in inter-organizational relationships (for example, between sites within a network, or between a practice and extramural physicians, other healthcare providers and laboratories)
- Changes in human capital (training of physicians, other healthcare providers and staff in performance measurement routines)
- Changes in financial capital (performance measurement costs).

### 2.1.2. Recommendation

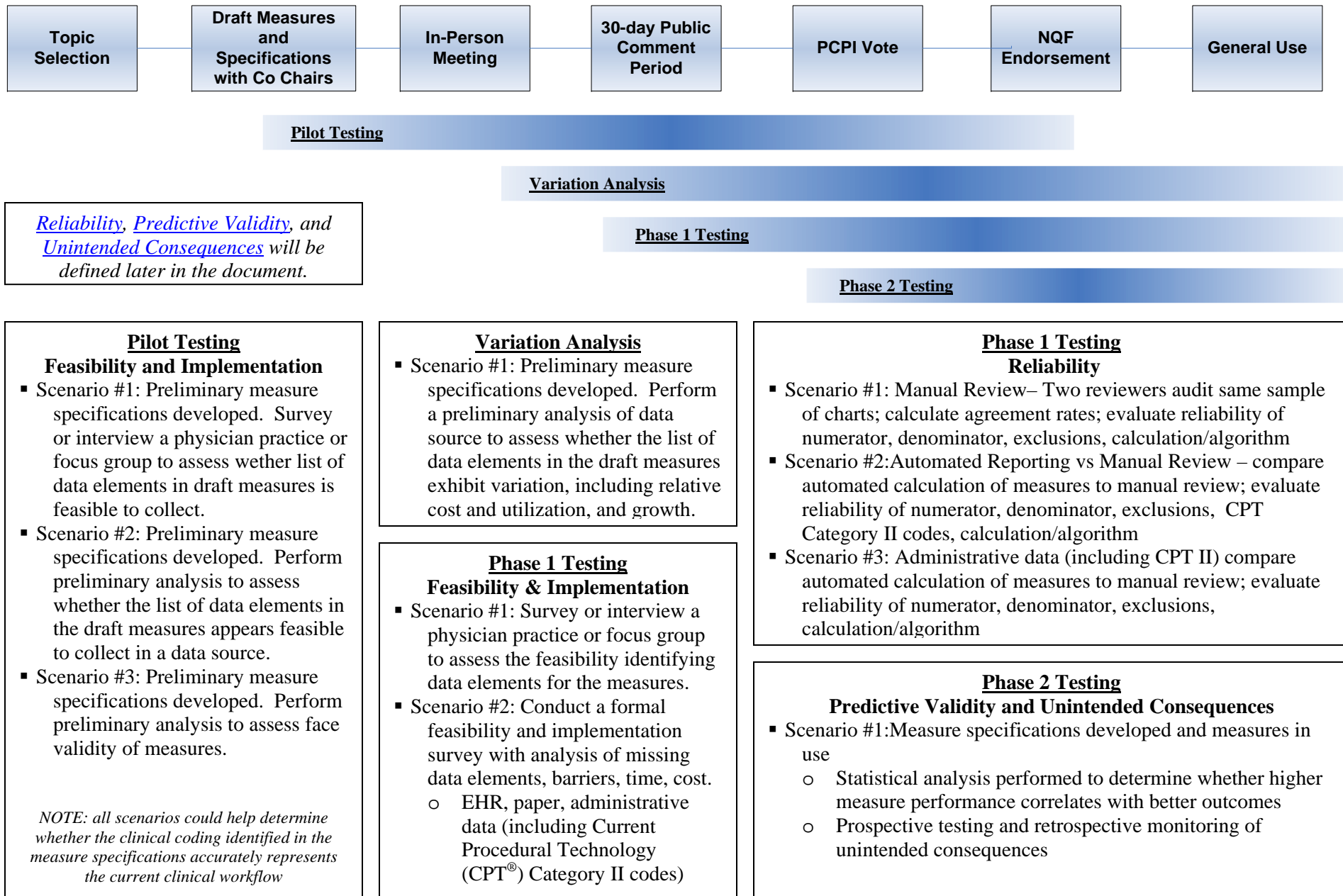
The PCPI MIE recommends that all PCPI performance measures be accompanied by at least one feasibility and implementation study that describes: the implementation strategy used to integrate performance measures and measurement within the organization of a practice; the feasibility of data collection (completeness of data collection, missing data problems); an analysis of barriers encountered in data collection, measure computation, and/or performance reporting; and an analysis of resource utilization and costs.

***Recommendation 2-1. At least one feasibility and implementation study should accompany all PCPI performance measures. Studies should include: (a) a description of the implementation strategy; (b) feasibility analysis of data collection; (c) barriers analysis; and (d) an analysis of resource utilization/costs.***

### 2.1.3. Rationale

Findings from feasibility and implementation studies may provide additional insights for PCPI measure developers to refine measure definition/specifications .

**Figure 1. Example of how Measure Testing can be Integrated into the Measure Development Cycle**



#### **2.1.4 Methodology – Pilot testing for Implementation**

There are two distinct and important ways for feasibility and implementation testing to inform measure development:

- a) One way is to test for feasibility and implementation after measure development, and use lessons learned in the measure maintenance update.
- b) A second way is to pilot test for feasibility and implementation during the measure development process and provide the lessons learned from testing to the measure development workgroup

Pilot testing, also referred to as “alpha testing” or “formative testing,” should be carried out during the initial measure development process. The measure development and testing timeline is shown in [Figure 1](#). The results are utilized to refine the draft specifications before they are finalized. Types of testing at this stage may vary significantly depending on the types of measures and data elements required. Typically, pilot testing is of a smaller scale than traditional testing, and the testing cycle may be carried out multiple times as revisions to the measures are made. Pilot testing will begin to assess the measure feasibility, barriers to implementation, and the burden of data collection and analysis.

The results of pilot testing will also inform the measure development process, as it is learned what factors make a measure more or less feasible or reliable. This feedback process is diagrammed in [Figure 2](#).

As a result of feasibility testing, it is important to show that practices can calculate performance rates. Preliminary **performance** results from testing concurrent with measure development can play a role in informing the measure development process. Performance results can provide valuable assistance with determining the size or existence of a variation in care or a practice gap, especially in areas that do not currently have a large literature base about the practice gap.

##### **2.1.4.1 Questionnaire Based Pilot Testing**

Some information about implementation can be gathered by sending questionnaires to practices. Sample questions are shown below, for *Pathology Measure #2: Colorectal Cancer Resection Pathology Reporting pT category (primary tumor) and pN category (regional lymph nodes) with histologic grade*.

- Is the diagnosis of cancer entered in a specific field in a routine way? Please describe this process:
- Is the histologic grade entered in a specific field in a routine way? Please describe this process:
- Are ICD-9 codes entered for the procedures? If not, is another coding system used?
- Is it feasible to develop and run a report from the computer-based system showing all patients within a certain timeframe who were diagnosed with Colorectal Cancer?
- Is it feasible to develop and run a report from the computer-based system showing all patients within a certain timeframe who have a Colorectal Cancer Resection Report? If yes, does this report already exist in your institution? If yes, it is possible to display on the report whether or not they met Measure #2?

- Do you currently monitor performance on these two pathology measures? If yes, who reviews the results? What decisions are made based on performance?

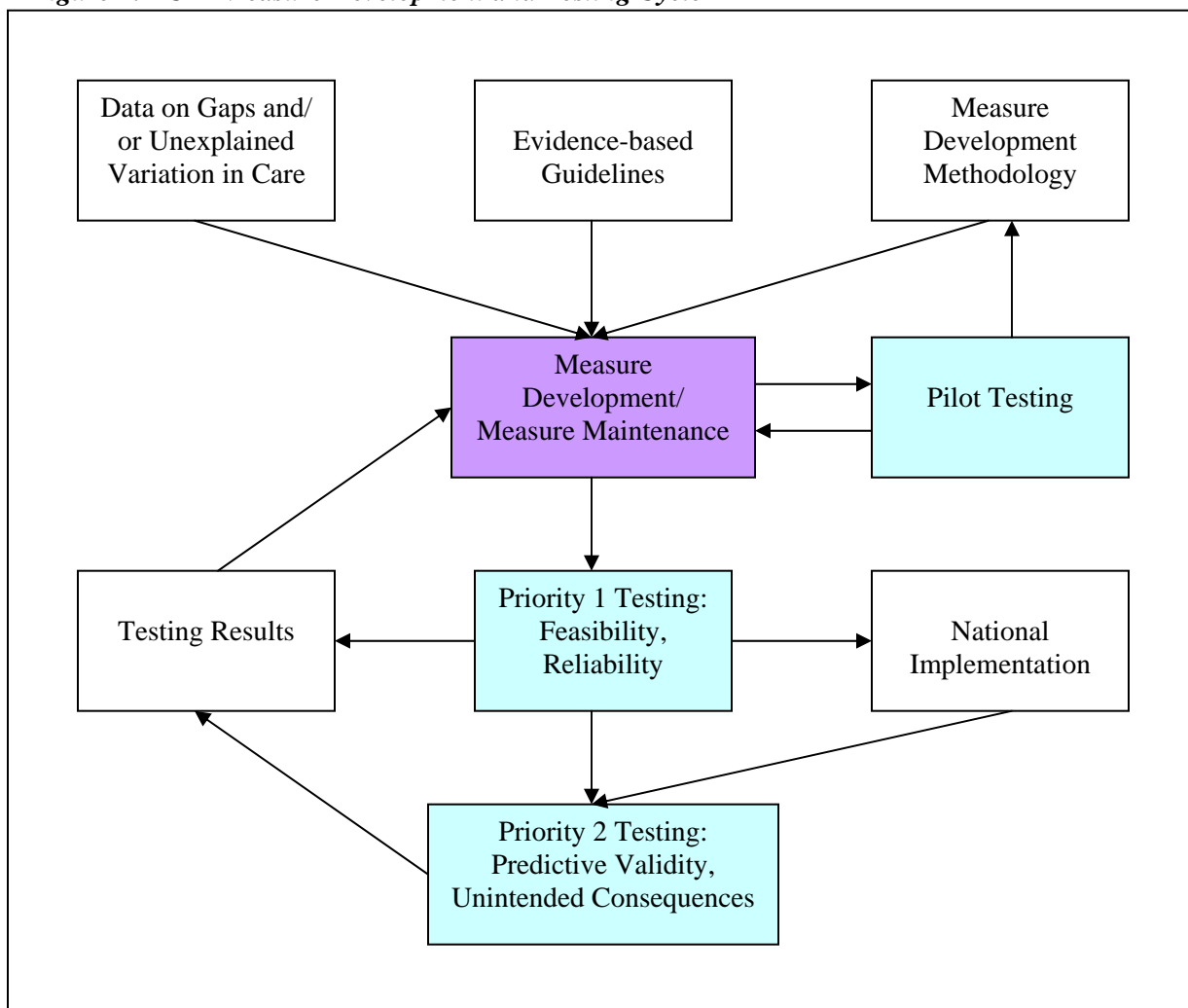
#### 2.1.4.2 Focus Group Based Pilot Testing

More in-depth information can be gathered by conducting a focus group at a potential implementation site. Sample questions are shown below, for *Care Transition #1: Reconciled Medication List Received by Discharged Patients*.

- Does your hospital require that a specific medication reconciliation form be used for all patients?
- If no, where in the patient’s record is medication reconciliation documented?
- Is the same form used for medical and surgical discharges?
- Is a medication reconciliation form required for pre-operative information?
- Who is responsible for medication reconciliation at your hospital?

It should be noted that pilot testing can also provide information about the face validity of measures. This is addressed further in section 4.2.4.

**Figure 2: PCPI Measure Development and Testing Cycle**



### **2.1.5. Methodology – Description of Implementation Strategy**

Project reports should include a narrative description of the implementation strategy.

#### **2.1.5.1. Practices using EHR-based measurement modalities**

**Electronic Health Record.** The International Standards Organization (ISO) defines an *electronic health record* as: “a repository of information regarding the health status of a subject of care, in computer processable form.”<sup>6</sup> For purposes of this Protocol we define an electronic health record (EHR) as an electronic form of a medical record or health record. The term electronic health record can refer to a single electronic medical record (for example, the medical record of a single patient stored in electronic format, multiple medical records in electronic format, or a computer software system which stores and provides user access to view and update medical records).

In practices employing *EHR-based measurement*, the narrative should describe the existing [EHR](#) product in use and its history of use within the practice (for example, the number of years the EHR has been use by the practice; general familiarity health professionals and staff have with the EHR product; and any previous experience with EHR-based performance measurement or quality improvement). The narrative must also include a description of: how a measure set’s definition and specifications were implemented and integrated within the EHR; how data are abstracted, cleaned, and verified for accuracy; how measure performance rates and exception rates are computed; and how reports are generated. The process of EHR-based performance measurement should be described with reference to a general activity timeline.

#### **2.1.5.2. Practices using administrative/claims data-based measurement modalities**

In practices employing *administrative/claims data-based measurement* (ie, measurement using administrative data), the narrative should describe the structure or format of existing administrative data and coding systems in use. The narrative should describe how specific data elements for the measure set under consideration are identified both in terms of the types of codes used (eg, ICD-9-CM, CPT Category II codes) as well as the location of these data in administrative records. Idiosyncrasies of data collection, such as fixed limits on the number of diagnostic and procedural codes that can be entered on a single claim and which may have implications for the accuracy of measurement, should be noted. The narrative should describe the abstractors and abstraction process, as well as data management, measure calculation, and performance reporting.

#### **2.1.5.3. Practices using paper medical records data-based measurement modalities**

In practices employing paper chart-based measurement, the narrative should describe the structure or format of paper charts, and how specific data elements for the performance measure under study are identified in the charts. The data abstraction form and process should be described, as should the processes of data management, measure calculation, and performance reporting.

#### **2.1.5.4. Practices using clinical registry modalities**

---

<sup>6</sup> International Standards Organization. *Health informatics – electronic health record – definition, scope, and context*. 2005. A draft version of this document: [http://www.openehr.org/downloads/isotc215wg3\\_N202\\_ISO-TR\\_20514\\_Final\\_%5B2005-01-31%5D.pdf](http://www.openehr.org/downloads/isotc215wg3_N202_ISO-TR_20514_Final_%5B2005-01-31%5D.pdf)

In practices employing clinical registries the narrative should describe the target patient population, the method of data entry, and the flow of data from the practice to the central warehouse/analysis agent. The methods for ensuring data quality and uniformity of data entry should be described. Registry-wide benchmarks for specific performance measures should be provided. For a given measure, reports should be provided to demonstrate the performance of a given practice compared to the benchmark.

#### **2.1.6. Methodology – Feasibility of Data Collection**

A standard feasibility analysis should quantify rates of missing data for each data element and for the measure as a whole. High rates of missing data may indicate: (1) potential mis-implementation or integration of performance measure definitions or specifications within the host health information system (ie, EHR system, administrative data-based abstraction system, or paper chart-based abstraction system); and/or (2) potential problems in the performance measure definitions and/or specifications. Both of these issues may provide useful information for refining measure definitions/specifications and for providing implementation guidance. Toward these ends, some effort beyond simple quantification of the frequency of missing data should be undertaken to ascertain the reason(s) why data are missing.

#### **2.1.7. Methodology – Barriers Analysis**

A standard feasibility and implementation study should enumerate and describe barriers encountered in: implementing/integrating performance measure definitions/specifications within the existing health information system; data abstraction; measure calculation; and performance reporting. Qualitative methods (for example, case studies, ethnographic observation, focus group assessments) and quantitative methods are both acceptable forms of research for barriers analysis.

Examples of barriers encountered during the implementation of the ACC/AHA/PCPI Beta-Blocker Performance Measure are presented in Box 2-1.

#### **Box 2-1. ACC/AHA/PCPI Beta-Blocker Performance Measure Examples of Implementation Barriers**

Barriers encountered in retrieving data elements for performance measures:

**Example from an [EHR](#) practice setting:** Data elements are not always in “searchable fields.” A physician may have prescribed beta blocker therapy for a patient, and indicated this action in the notes section, but not included “beta blocker” on the searchable medication list.

**Example from a paper chart practice setting:** We have learned from previous pilot testing experiences that for the beta blocker measure, historical elements, such as prior myocardial infarction and whether a patient discontinued beta blocker therapy for a specific reason in the past, are often difficult to locate in the chart. Also, there may be conflicting histories noted among the primary care and consulting physicians.

From the perspective of the PCPI, a thorough assessment of barriers as well as their potential causes and possible solutions, is critical information that can feed back into the overall cycle of performance measure development. Problems that may point to issues in specification can be directed back to the measure developers for further technical refinement of the performance

measure being tested. This information may also be relevant to concurrent measure development efforts within PCPI and provide guidance for measure development activities beyond the scope of a single measure.

### **2.1.8. Methodology – Resource Utilization/Cost Analysis**

The objective of the resource utilization and cost analysis is to estimate the costs attributable to implementation, measurement, and reporting for the specific measure or measure set. That is, in the evaluation of a specific PCPI measure or measure set, the PCPI is interested in the incremental costs associated with implementing the measure/measure set. If a clinical practice already engages in performance measurement, the incremental resources/costs of implementing a new measure set will be a subset of overall resources expended in performance measurement and/or quality improvement.

Investigators should explicitly present their cost or accounting model. [Figure 3](#) provides one hypothetical example of a general framework for identifying resource use and costs that is adapted from the OECD Standard Cost Model<sup>7</sup>. A specific performance measure set under study is comprised of one or more individual performance measures (1,...,  $M$  performance measures). Each individual performance measure imposes four discrete information obligations: denominator; numerator; denominator exception; and the performance measure itself. Each information obligation requires the collection and processing of one or more data elements (1,...,  $D$  data elements). These data elements are the discrete data elements identified in the technical specifications accompanying a specific measure set. The acquisition and processing of each data element to fulfill each information obligation involves one or more discrete activities. For example:

- in an [EHR](#)-based measurement system, activities may include programming definition and specifications for an individual measure for automated abstraction of data elements, data cleaning and data verification;
- in an administrative data or paper chart-based measurement system, activities may include the construction of abstraction forms; abstractor training; abstraction; data compilation, data cleaning, and data verification.

Activities associated with measure implementation include calculating the performance measure and reporting the measure. Each activity is associated with costs.

- *Internal* costs include the hourly wage rate of internal staff working on the specific measure set and overhead costs.
- *External* costs are those incurred if a practice outsources performance management activities to an extramural organization or business.
- *Acquisition* costs are those associated with resources that were acquired and consumed exclusively in the process of implementing, measuring, and reporting a specific measure set.

General considerations to guide the enumeration of resources and costs consumed in implementation, measurement, and reporting for a specific measure set are as follow.

- Capital and other resources used for multiple purposes beyond the implementation, measurement, and reporting of the measure set under study should not be counted as a direct cost. Rather, these costs should be subsumed as overhead.

---

<sup>7</sup> Modeled after: International Standard Cost Model. Standard Cost Model Network. <http://www.oecd.org/dataoecd/32/54/34227698.pdf>

- Overhead costs will vary across practices. Cost analyses must make explicit what is subsumed under overhead, and should be explicit about assumptions and discount rates. Justifications for assumptions should be provided.
- Acquisitions and resources to be included as direct costs are those that are essential to the implementation, measurement, and reporting of a specific performance measure set under study, and which are used exclusively for the purpose of implementing, measuring, or reporting a specific measure set.
- The time frame for reporting costs is to be a financial/accounting year. If an acquisition or resource has a functional lifetime exceeding one year, than an approximation should be used: the total cost of the acquisition/resource should be divided by the expected lifetime (in years).
- It would be useful for analysts to distinguish between *one-time* costs and *recurring* costs. One-time costs are costs that are incurred by a practice in implementing a specific measure set for the first time. Recurring costs are those incurred by a practice each time measurement and reporting are undertaken for a specific measure set.

## 2.2. Scope of Feasibility and Implementation Testing

### 2.2.1. Recommendation

To ensure thorough evaluation of the PCPI measures and specifications, all PCPI measures should be tested in various settings, including but not limited to: clinical practices utilizing [EHR](#)-based performance measurement; clinical practices utilizing administrative data-based performance measurement; and clinical practices utilizing paper medical chart-based performance measurement. Of particular interest is the investigation of potential biases which may be inherent in the use of a particular technical specification in particular practice settings. Measurement activities that may be feasible in an EHR-enabled environment may not be feasible within reasonable margins of time and cost in administrative data-based and paper chart-based measurement environment. It is important that each specification for each existing PCPI measure functions in a manner that provides a level playing field for measurement and reporting by a variety of practice types. Because EHR adoption has largely occurred among larger, more affluent practices, feasibility testing in EHR-, administrative data-, and paper chart-based measurement environments will also help ensure testing in practice settings of various sizes, as smaller practices are currently more likely to be characterized by administrative- or paper chart-based performance measurement.

***Recommendation 2-2. Feasibility of implementing PCPI performance measures should be demonstrated in the following settings: (a) clinical practices using EHR-based measurement modalities; (b) clinical practices using administrative/claims data-based measurement modalities; and (c) clinical practices using paper medical record data-based measurement modalities.***

### 2.2.2. Rationale

Feasibility is not an intrinsic property of a performance measure set; rather, it depends on characteristics of the practice environment where a measure set is implemented, and on the

manner in which it is integrated within an organization. The feasibility of an existing PCPI measure may vary by practice settings as a result of infrastructural and organizational differences.

The research recommendations regarding the scope of feasibility and implementation testing are intended to ensure that the feasibility and implementation of existing PCPI measures are assessed in ambulatory practice settings across each of the three measurement modalities that the measures have been developed for: [EHR](#)-based, administrative/claims data-based, and paper medical record data-based measurement. In addition, PCPI measures should be tested in safety-net settings, such as federally-qualified health centers, to assess feasibility and implementation under resource constraints and unique environmental demands.

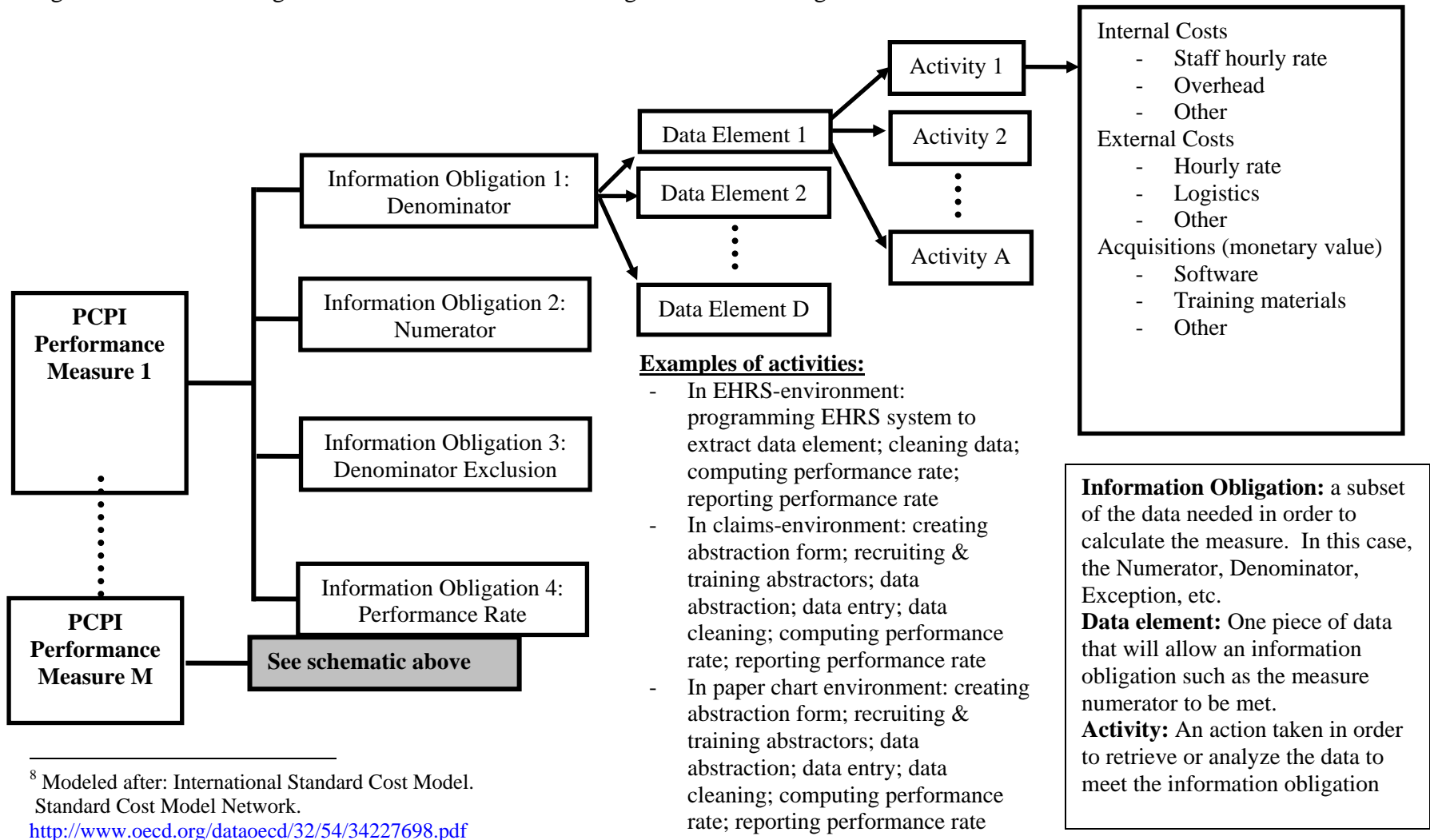
In the case of EHR-based measurement modalities, we recognize that, at this time, EHR products are heterogeneous. Demonstrating feasibility and implementation using one EHR product does not imply that measurement and reporting is equally feasible to implement across all EHR products currently in use. Despite this limitation, it is not possible to require each performance measure to be demonstrated feasible in all major EHR systems. Recommendation 2-2 will be considered satisfied if the feasibility and implementation of a measure has been assessed in at least one study in the context of one EHR product.

### 2.2.3 Recommendation

***Recommendation 2-3. Feasibility of implementing PCPI performance measures should be demonstrated in private and public clinical environments.***

Practices serving disproportionate shares of clinically and/or sociodemographically vulnerable patients face different environmental, organizational, and clinical constraints that may have different implications for the feasibility of a performance measure set compared to other private practices.

Figure 3. Schematic diagram of a cost model for estimating the costs of a single a PCPI Measure Set<sup>8</sup>



<sup>8</sup> Modeled after: International Standard Cost Model. Standard Cost Model Network. <http://www.oecd.org/dataoecd/32/54/34227698.pdf>

**Exhibit 1. Example Table to Show Measure Performance Results, With and Without Exceptions**

| <i>MEASURE</i>     | <i>MEASURE RATE WITHOUT EXCEPTIONS</i> | <i>MEASURE RATE WITH EXCEPTIONS*</i> | <i>EXCEPTION RATE*</i> |
|--------------------|--|--------------------------------------|------------------------|
| <b>Measure Set</b> |  |                                      |                        |
| <i>Measure 1</i>   | %                                      | %                                    | %                      |
| <i>Measure 2</i>   | %                                      | %                                    | %                      |
| <i>Measure 3</i>   | %                                      | %                                    | %                      |

\* If a measure were to have no exceptions, mark the row as “This measure is not specified with exceptions”

**Exhibit 2. Example Table to Show Measure Exceptions Documentation**

| <i>Measure</i>   | <i>Exception Type</i> | <i>Verbatim Documentation For Exceptions*</i> |
|------------------|-----------------------|---|
| <i>Measure 1</i> |                       |   |
|                  | Medical               | Comfort care only                             |
|                  | Medical               | Patient is ambulatory                         |
| <i>Measure 2</i> |                       |   |
|                  | Medical               | Medication not ordered due to bleeding issues |
|                  | Medical               | Thrombocytosis                                |
| <i>Measure 3</i> |                       |   |
|                  | Medical               | Allergic to medication                        |
|                  | Medical               | Intracerebral bleed                           |
|                  | Patient               | Comfort care only due to brain hemorrhage     |
|                  | Patient               | Patient refuses medication                    |

- If a measure were to have no exceptions, mark the row as “This measure is not specified with exceptions”

**Exhibit 3. Example Table to Show Percentage Codified Data and Percentage Discrete Data**

| <i>Measure</i>         | <i>% Discrete Data</i> | <i>% Codified Data</i> |
|------------------------|------------------------|------------------------|
| <i>Measure 1</i>       |                        |                        |
| <i>Data element 1A</i> | %                      | %                      |
| <i>Data element 1B</i> | %                      | %                      |
| <i>Measure 2</i>       |                        |                        |
| <i>Data element 2A</i> | %                      | %                      |
| <i>Data element 2B</i> | %                      | %                      |

|                        |   |   |
|------------------------|---|---|
| <i>Measure 3</i>       |   |   |
| <i>Data element 3A</i> | % | % |
| <i>Data element 3B</i> | % | % |
| <i>Data element 3C</i> | % | % |
| <i>Data element 3D</i> | % | % |

**Exhibit 4. Example Table to Show Cost Analysis**

|                                    |           |
|------------------------------------|-----------|
| Average Abstraction Time per chart | x minutes |
| Hourly Rate for abstractor         | \$x.00    |
| Estimated Total Abstraction Cost   | \$x.00    |
| Cost of Abstraction per Record     | \$x.00    |
| Time for Abstraction per Record    | x minutes |

## SECTION III. TESTING AREA 3: RELIABILITY

### Section III Outline

- 3.1. [Reliability and PCPI Performance Measures](#)
  - 3.1.1. [Definition](#)
  - 3.1.2. [Recommendation](#)
  - 3.1.3. [Rationale](#)
  - Table 3-1. [Recommendation 3-1 Verification Tests](#)
- 3.2. [Inter-rater \(Inter-abstractor\) Reliability](#)
  - 3.2.1. [Definition](#)
  - 3.2.2. [Recommendation](#)
  - 3.2.3. [Rationale](#)
  - 3.2.4. [Methodology](#)
- 3.3. [Parallel Forms Reliability](#)
  - 3.3.1. [Definition](#)
  - 3.3.2. [Recommendation](#)
  - 3.3.3. [Rationale](#)
  - 3.3.4. [Methodology \(Parallel Forms Reliability\): EHR-based Measurement Modalities vs. Manual Review](#)
  - 3.3.5. [Methodology \(Parallel Forms Reliability\): Administrative/Claims Data-based Measurement Modalities vs. Manual Review](#)
- 3.4. [Other Forms of Reliability Testing](#)
  - 3.4.1. [Test-retest Reliability](#)
  - 3.4.2. [Internal Consistency](#)
- 3.5. [Scope of Reliability Testing](#)
  - 3.5.1. [Recommendation](#)
  - 3.5.2. [Remarks](#)

### 3.1. Reliability and PCPI Performance Measures

#### 3.1.1. Definition

*Reliability* refers to “the stability of a set of observations generated by an indicator under a fixed set of conditions, regardless of who collects the observations or of when or where they are collected,”<sup>9</sup> and is a scientific attribute of measurement instruments (for example, assessment surveys, screening questionnaires, screening tests, competency tests, assays, or technical instruments).

#### 3.1.2. Recommendation

The PCPI MIE recommends that all PCPI performance measures undergo reliability testing in all measurement modalities for which technical specifications are developed. Currently, these measurement modalities are: EHR- or registry-based measurement, administrative/claims-based measurement, and paper medical record data-based measurement.

***Recommendation 3-1. All PCPI performance measures should undergo reliability testing in all measurement modalities for which technical specifications are developed: EHR- or registry- based measurement, administrative/claims-based measurement, registry–based measurement and paper medical record data-based measurement.***

#### 3.1.3. Rationale

Ideally, measurement of a specific physician’s performance based on a set of patients should be the same regardless of whether data collection, case finding, and performance measure calculations are accomplished through implementation of EHR- or registry-based measurement strategy, administrative/claims-based measurement strategy, or a paper medical record data-based measurement strategy. Performance measures and the technical specifications developed for each performance measure should yield stable, consistent measurements across measurement modalities. To the extent that the definitions and technical specifications developed by PCPI enable users to obtain the same measurements of a physician’s performance across multiple measurement modalities, it is said that a performance measure and its specifications have demonstrated reliability across performance measurement modalities.

Reliability across performance measurement modalities is an important attribute because reporting practices will vary in their capacity to use and report PCPI measures. Some practices will be able to use EHR-based strategies. Other practices that do not have an EHR in place and who do not have resources to conduct reviews of paper charts, may use administrative/claims-based measurement and reporting strategies. In order for performance measures to yield informative data for decision-making and quality improvement, a measure and its specifications should provide stable measurements of performance independent of the technology and process used to make those measurements.

All PCPI performance measures are developed with definitions, technical specifications for EHR-based measurement, administrative/claims data-based measurement, as well as paper medical

---

<sup>9</sup> Cohen BP. *Developing Sociological Knowledge*. Chicago: Nelson-Hall, 1989. Page 155.

record data-based measurement to broaden the use of the measures by all types of practices. It is thus recommended that all PCPI measures be evaluated for reliability across the various measurement modalities.

The reliability of PCPI measures across different measurement modalities may be evaluated by:

- Assessing inter-abstractor reliability in denominator, numerator, and exception case finding as well as the calculation of whole measures in paper or electronic medical record chart-based measurement strategies – this is basically “inter-rater” or “inter-abstractor” reliability and is discussed in [Section 3.2.](#), below
- Verifying EHR- based strategies for identifying denominators, numerators, and exceptions as well as the calculation of whole measures against manual review of patient records (EHR- or registry-based measurement modality vs. manual review of records) – this is a type of “parallel forms reliability” and is discussed in [Section 3.3.](#), below
- Verifying administrative/claims-based strategies for identifying denominators, numerators, and exceptions as well as the calculation of whole measures against manual review of patient records (administrative/claims-based measurement modality vs. manual review of records) – this is a type of “parallel forms reliability” and is discussed in [Section 3.3.](#), below
- Verifying registry- or electronic data warehouse-based strategies for identifying denominators, numerators, and exceptions as well as the calculation of whole measures against manual review of patient records (registry-based measurement modality vs. manual review of records) – this is a type of “parallel forms reliability” and is discussed in [Section 3.3.](#), below

Table 3-1 provides a map of verification tests that are implied in Recommendation 3-1, and which collectively constitute tests of the reliability of a performance measure across measurement modalities.

**Table 3-1. Recommendation 3-1 Verification Tests**

|                                 | <b>EHR-Based Measurement</b>                                   | <b>Administrative/ Claims-Based Measurement</b>                | <b>Registry-Based Measurement</b>                              | <b>Medical Record- Based Measurement</b>         |
|---------------------------------|--|--|--|--|
| <b>DENOMINATOR Case-finding</b> | <b>1A</b><br>vs. Manual Review<br>(Parallel Forms Reliability) | <b>2A</b><br>vs. Manual Review<br>(Parallel Forms Reliability) | <b>3A</b><br>vs. Manual Review<br>(Parallel Forms Reliability) | <b>4A</b><br>Inter-abstractor Reliability        |
| <b>NUMERATOR Case-finding</b>   | <b>1B</b><br>vs. Manual Review<br>(Parallel Forms Reliability) | <b>2B</b><br>vs. Manual Review<br>(Parallel Forms Reliability) | <b>3B</b><br>vs. Manual Review<br>(Parallel Forms Reliability) | <b>4B</b><br>Inter-abstractor Reliability        |
| <b>EXCUSION Case-finding</b>    | <b>1C</b><br>vs. Manual Review<br>(Parallel Forms Reliability) | <b>2C</b><br>vs. Manual Review<br>(Parallel Forms Reliability) | <b>3C</b><br>vs. Manual Review<br>(Parallel Forms Reliability) | <b>4C</b><br><b>Inter-abstractor Reliability</b> |
| <b>OVERALL Measure</b>          | <b>1D</b>  | <b>2D</b>  | <b>3D</b>  | <b>4D</b>  |

## **3.2. Inter-rater (inter-abstractor) reliability**

### **3.2.1. Definition**

*Inter-rater reliability* refers to the extent to which observations from two or more human observers are congruent with each other. When a measurement procedure or instrument relies on human observation or judgment, the procedure or instrument should be specified and constructed in a manner that facilitates uniform observation to the extent possible. In the context of this testing Protocol, inter-abstractor reliability is a synonym for inter-rater reliability.

### **3.2.2. Recommendation**

The PCPI MIE recommends that evaluations of inter-abstractor reliability be conducted for all PCPI performance measures.

**Recommendation 3-2. All PCPI performance measures should undergo inter-abstractor reliability testing to evaluate medical record data-based measurement modalities.**

### **3.2.3. Rationale**

In the context of PCPI performance measures, inter-rater reliability is applicable primarily in the context of medical record-based abstraction (whether electronic- or paper- based), in which human chart abstractors must review charts and extract data using abstraction tools developed by PCPI. In EHR-based and administrative/claims data-based performance measurement, parallel forms reliability should also be assessed whenever possible (see 3.3.).

### **3.2.4. Methodology (Inter-abstractor reliability): Medical record data-based Measurement Modalities**

#### **3.2.4.1. Terms**

*Performance measure patient population (Population).* The performance measure patient population (Population) is the set of patients during a measurement period (testing period, audit period, or reporting period) over which performance measures will be calculated. The Population is a subset of a physician's (or practice's) entire case base extracted on the basis of broad performance eligibility criteria. For example, in the PCPI/American College of Cardiology (ACC)/American Heart Association (AHA) Coronary Artery Disease (CAD) measure set, the eligibility criteria that define the relevant Population for calculating individual CAD performance measures are: (a) all patients who had at least two face-to-face office visits with the physician; (b) all patients aged 18 years and older at the beginning of the measurement time period; AND (c) all patients who have a documented diagnosis of coronary artery disease.

*Performance measure patient population sample (Population Sample).* A performance measure patient population sample (Population sample) is a subset of patients in the Population (see above).

*Manual denominator/numerator/exception case finding strategy (Manual Case Finding Strategy).* The manual medical records-based denominator/numerator/exception case finding strategy refers

to the manner in which a physician practice translates the PCPI measure definition and develops procedures for implementing the chart abstraction tools developed by PCPI to:

- Extract data elements necessary for case identification and performance measure calculation from an administrative or claims (billing) database;
- Identify cases (determine which cases are to be included in the denominator/numerator/exception set) by using the PCPI abstraction tools and algorithms; and
- Performance measure calculation.

In practice, the *Manual Case Finding Strategy* will include such procedures as: training staff in using PCPI chart abstraction tools and abstraction protocols and process; data entry and processing; and measure computation.

### 3.2.4.2. Assessing the reliability of manual denominator case finding strategies (3A)

The *Manual Case Finding Strategy* should be applied to the performance measure patient population (Population) or a sample from the Population to identify cases to be included in the denominator by one or more abstractors. All patients in the Population or Population sample should be classified by the Manual Case Finding Strategy (Manual CFS) as belonging in the denominator (Manual CFS1 DEN+) or not belonging in the denominator (Manual CFS1 DEN-).

The Manual CFS should be applied to the same Population or Population sample by another abstractor or team of abstractors. All patients in the Population or Population sample should be classified by the second abstractor using the Manual CFS as either belonging in the denominator (Manual CFS2 DEN+) or not belonging in the denominator (Manual CFS2 DEN-).

Cross-classification of patients in the Population or Population sample by denominator inclusion status as determined by the two abstractors (or abstraction teams) using the same Manual CFS yields the following 2x2 contingency table (Table 3.2.4.2).

**Table 3.2.4.2. Assessing the reliability of medical record-based denominator case finding**

|                   | Manual CFS 2 DEN+          | Manual CFS 2 DEN-          |                   |
|-------------------|----------------------------|----------------------------|-------------------|
| Manual CFS 1 DEN+ | Inter-abstractor AGREEMENT | DISAGREEMENT               |                   |
| Manual CFS 1 DEN- | DISAGREEMENT               | Inter-abstractor AGREEMENT | Total: Population |

The extent of inter-abstractor reliability – i.e. agreement on cases to be included in the denominator – should be quantitatively summarized. Concordance rates and Cohen’s Kappa with confidence intervals are acceptable statistics to describe inter-abstractor reliability in performance measurement based on paper medical records.

### 3.2.4.3. Assessing the reliability of medical records-based numerator case finding strategies (3B)

The Manual Case Finding Strategy ([Manual CFS](#)) should be applied to the identified patients in the denominator identified by Manual CFS to identify cases for inclusion in the numerator. All patients in the denominator should be classified by the first abstractor or team of abstractors using

the Manual CFS as belonging in the numerator (Manual CFS1 NUM+) or not belonging in the numerator (Manual CFS1 NUM-)

The Manual CFS should be applied to the denominator identified by Manual CFS by another abstractor or team of abstractors. All patients in the denominator should be classified by the second group of abstractors using the Manual CFS as either belonging in the numerator (Manual CFS2 NUM+) or not belonging in the numerator (Manual CFS2 NUM-).

Cross-classification of patients in the denominator by numerator inclusion status as determined by the two abstractors (or abstraction teams) using the same Manual CFS yields the following 2x2 contingency table (Table 3.2.4.3).

It should be noted that it may be necessary to adjudicate any discrepancies between the two abstractors, in the Manual CFS identified denominator, before applying the Manual CFS numerator strategy.

**Table 3.2.4.3. Assessing the reliability of paper medical record-based numerator case finding**

|                         |                            |                            |  |
|-------------------------|----------------------------|----------------------------|--|
|                         | <b>Manual CFS2 NUM+</b>    | <b>Manual CFS2 NUM-</b>    |  |
| <b>Manual CFS1 NUM+</b> | Inter-abstractor AGREEMENT | DISAGREEMENT               |  |
| <b>Manual CFS1 NUM-</b> | DISAGREEMENT               | Inter-abstractor AGREEMENT | Total: Manual CFS identified denominator |

The extent of inter-abstractor reliability – i.e. agreement on cases to be included in the numerator – should be quantitatively summarized. Concordance rates and Cohen’s Kappa with confidence intervals are acceptable statistics to describe inter-abstractor reliability in performance measurement based on paper medical records.

**3.2.4.4. Verifying Manual exception case finding strategies**

The medical record-based case exception finding strategy (Manual CFS1 EXC) should be applied to all cases in the Manual CFS -identified denominator that were not identified for inclusion in the numerator – i.e. “apparent quality failures” ((Manual CFS DEN+)-( Manual CFS NUM+)). All apparent quality failures should be identified by Manual CFS1 EXC as either belonging in the excluded set (Manual CFS1 EXC+) or not belonging in the excluded set (Manual CFS1 EXC-).

This strategy will only identify exceptions for patients who have not met the measure. These exceptions are referred to as “applied exceptions,” as the physician has chosen not perform the process indicated by the measure due to the exception. In some study protocols, it is interesting and useful to instead identify exceptions for all patients who are eligible for the denominator. In the case of patients who have met the measure but who also have exceptions these exceptions are referred to as “not-applied exceptions,” as the physician has used his or her judgment to perform the process despite the exception.

The Manual CFS EXC should be applied to the same apparent quality failures by another abstractor or team of abstractors (Manual CFS2 EXC). All apparent quality failures should be classified by Manual CFS2 EXC as either belonging in the excluded set (Manual CFS2 EXC+) or not belonging in the excluded set (Manual CFS2 EXC-).

Cross-classification of apparent quality failures by exception status as determined by the two abstractors (or abstraction teams) using the same Manual CFS yields the following 2x2 contingency table (Table 3.2.4.4).

**Table 3.2.4.4. Assessing the reliability of medical record-based exception case finding**

|                         |                            |                            |  |
|-------------------------|----------------------------|----------------------------|--|
|                         | <b>Manual CFS2 EXC+</b>    | <b>Manual CFS2 EXC-</b>    |  |
| <b>Manual CFS1 EXC+</b> | Inter-abstractor AGREEMENT | DISAGREEMENT               |  |
| <b>Manual CFS1 EXC-</b> | DISAGREEMENT               | Inter-abstractor AGREEMENT | Total: Manual CFS identified apparent quality failures |

The extent of inter-abstractor reliability – i.e. agreement on cases to be excluded – should be quantitatively summarized. Agreement rates and Cohen’s Kappa with confidence intervals are acceptable statistics to describe inter-abstractor reliability in performance measurement based on medical records. Kappa statistics are preferred where possible as they take into account the agreement as compared to what is expected by chance.

**3.2.4.5. Assessing the reliability of overall performance measurement using medical record-based strategies (3D)**

In addition to assessing the reliability of measure components derived from medical record-based measurement strategies, the reliability of overall performance measures should also be assessed across different abstractors. The purpose of this test is to assess the distortion in overall performance calculations due to inaccuracies in Manual CFS. Table 3.2.4.5 provides an overview of the calculations and comparisons involved in evaluating the inter-abstractor reliability of overall performance measures.

**Table 3.2.4.5. Assessing the reliability of overall medical record-based performance measures**

|                        | <b>Medical Record – Abstractor 1</b>  | <b>Medical Record – Abstractor 2</b>  | <b>Difference</b>         |
|------------------------|---|---|---------------------------|
| <b>DENOMINATOR</b>     | Apply <a href="#">Manual CFS</a> (denominator) to SAMPLE to identify denominator<br>$DEN_{MR1}$ | Apply Manual CFS (denominator) to SAMPLE to identify denominator<br>$DEN_{MR2}$                 | $DEN_{MR1} - DEN_{MR2}$   |
| <b>NUMERATOR</b>       | Apply Manual CFS (numerator) to $DEN_{MR1}$ to identify numerator<br>$NUM_{MR1}$                | Apply Manual CFS (numerator) to $DEN_{MR2}$ to identify numerator<br>$NUM_{MR2}$                | $NUM_{MR1} - NUM_{MR2}$   |
| <b>EXCUSION</b>        | Apply Manual CFS (exception) to $(DEN_{MR1} - NUM_{MR1})$ to identify exceptions<br>$EXC_{MR1}$ | Apply Manual CFS (exception) to $(DEN_{MR2} - NUM_{MR2})$ to identify exceptions<br>$EXC_{MR2}$ | $EXC_{MR1} - EXC_{MR2}$   |
| <b>OVERALL Measure</b> | $PERF_{MR1} = NUM_{MR1} / (DEN_{MR1} - EXC_{MR1})$  | $PERF_{MR2} = NUM_{MR2} / (DEN_{MR2} - EXC_{MR2})$  | $PERF_{MR1} - PERF_{MR2}$ |

Performance measures should be calculated independently using data from abstractor 1 and abstractor 2. Independently, performance measures should be calculated using denominators,

numerators, and exceptions identified through manual review as applied to all cases in the Population or Population sample.

Formal statistical tests for differences in proportions, such as the Z-test, could be performed. Reasons for discrepancies between administrative/claims data-based components/measures and components/measures derived from manual review should be investigated and documented.

#### **3.2.4.6. Process Evaluation**

Low concordance rates, correlation coefficients, and/or low Kappa scores may indicate problems with PCPI abstraction tools (e.g., poor structure, insufficient definitions), but they may also indicate problems in the measurement strategy that are beyond the control of measure developers, such as:

- Inadequate abstractor training,
- Inadequate abstractor experience or competence, or
- Mis-matched abstractor experience.

Thus, in tests evaluating the reliability of paper medical record-based measurement, it is essential that investigators thoroughly document the following:

- Selection of abstractors/abstractor qualifications and experience;
- Abstractor training;
  - Trainer
  - Duration of training
  - Assessment of competence implementing the paper medical record based measurement strategy
  - Inter-abstractor reliability achieved after training, prior to formal review of records
  - Procedures for adjudication between abstractor discrepancies
- Abstraction procedures;
  - Use of PCPI abstraction tools
  - Any modifications or elaborations to PCPI abstraction tools
  - Abstraction definitions and instructions
- Evaluation of the inter-abstractor reliability of individual data elements retrieved. Forms that may be used to facilitate the reporting of the quality of individual data elements may be found in the exhibits at the end of this section. These forms provide a structured means for reporting: discrepancies in individual data elements abstracted by different abstractors; missing data; barriers encountered in retrieving data elements through paper-based methods involving manual abstraction.

PCPI will review methodological documentation of abstraction procedures in assessing the validity of test results.

### **3.3. Parallel forms reliability**

#### **3.3.1. Definition**

*Parallel forms reliability* assesses the extent to which multiple formats or versions of a test yield the same results.

### **3.3.2.Recommendation**

For all PCPI performance measures, the PCPI MIE recommends that EHR-based, administrative/claims data-based, and registry-based measurement modalities be verified against manual review. These verification tests are a form of parallel forms reliability that involve verifying one form of measurement – EHR- or administrative/claims data- or registry-based measurement – against a second form of measurement – manual review of the medical record.

**Recommendation 3-3. All PCPI performance measures should undergo the following tests of parallel forms reliability: (a) EHR-based measurement vs. manual review; (b) administrative/claims-based measurement vs. manual review; (c) registry-based measurement vs. manual review.**

### **3.3.3. Rationale**

In the context of PCPI performance measures, parallel forms reliability may be taken to refer to the extent to which different technical specifications (for EHR, administrative/claims, registry, or paper charts) accompanying a performance measure all yield the same results when applied to the same observations.

### **3.3.4. Methodology (Parallel Forms Reliability): EHR- or Registry-based Measurement Modalities vs. Manual Review**

#### **3.3.4.1. Terms**

*Performance measure patient population (Population).* The performance measure patient population (Population) is the set of patients during a measurement period (testing period, audit period, or reporting period) over which performance measures will be calculated. The Population is a subset of a physician (or practice) total case base extracted on the basis of broad performance eligibility criteria. For example, in the PCPI/ACC/AHA CAD measure set, the eligibility criteria that define the relevant Population for calculating individual CAD performance measures are: (a) all patients who had at least two face-to-face office visits with the physician; (b) all patients aged 18 years and older at the beginning of the measurement time period; AND (c) all patients who have a documented diagnosis of coronary artery disease.

*Performance measure patient population sample (Population Sample).* A performance measure patient population sample (Population Sample) is a subset of the patients in the performance measure patient population (Population). The population sample should be identified by a scientifically acceptable method, such as random sampling, or selecting all cases within a certain timeframe such as a year.

*Automated denominator/numerator/exception case finding strategy (Automated CFS).* The automated denominator/numerator/exception case finding strategy (Automated CFS) refers to the

manner in which a physician practice translates the PCPI measure definition and technical specifications into a procedure for:

- Extracting data elements necessary for case identification and performance measure calculation from the electronic health record (EHR) or registry
- The process of case identification: (determining which cases are to be included in the denominator/numerator/exception set by applying selection algorithms/rules specified by PCPI to data from an EHR or registry)
- Performance measure calculation

In practices using EHR- or registry-based measurement, a case-finding strategy will refer generally to the manner in which PCPI measure definitions and technical specifications are translated into automated information retrieval and case identification procedures within a local EHR or registry. PCPI currently provides a single version of technical specifications for electronic health records. However, EHR and registry products vary substantially, and even the same product may be used in very different ways by different practices.

*Reference Strategy (also known as a gold standard).* Verification of data elements obtained through automated search strategies of electronic health records, registries, or administrative claims data should be conducted against some reference strategy for obtaining the data elements. It is common practice in the empirical literature to use manual review of data as the reference strategy against which automated data search and extraction strategies are evaluated. In studies verifying data elements extracted using automated search and extraction procedures from electronic health record data, data elements retrieved from automated strategies are typically compared against manual review of the electronic health record.<sup>10</sup> In studies verifying data elements extracted using automated procedures from electronic claims databases, data elements retrieved from automated strategies are often compared against manual review of administrative claims.<sup>11</sup> Manual review is not perfect.<sup>12</sup> However, in the context of this Protocol, for the purposes of verifying data elements to test the reliability of PCPI measures, we recommend that automated data search and extraction procedures be compared against manual review. In this sense, we regard manual review -- whether of health records (electronic or paper) as the reference strategy against which automated search and extraction of data should be compared. We regard data obtained by manual review (whether of medical records (electronic or paper), as what is typically referred to as a gold standard in the specific context and purpose of this document. We will henceforth use the phrases, *reference strategy* and *gold standard* interchangeably. Please note: we are not upholding electronic data (whether electronic health records or electronic administrative data) as gold standards – we are only upholding manual review as a reference strategy for obtaining data from data sources.

---

<sup>10</sup> For examples, see: Persell SD, Wright JM, Thompson JA, Kmetik KS, Baker DW. Quality measures for coronary artery disease using an electronic health record. *Arch Intern Med.* 2006;166:2272-2277. Baker DW, Persell SD, Thompson JA, Soman NS, Burgner KM, Liss D, Kmetik KS. Automated review of electronic health records to assess quality of care for outpatients with heart failure. *Ann Intern Med.* 2007;146:270-277.

<sup>10</sup> Peabody JW, Luck J, Jain S, Bertenthal D, Glassman P. Assessing the accuracy of administrative data in health information systems. *Med Care.* 2004;42(11):1066-72. Also: Kashner TM. Agreement between administrative files and written medical records: a case of the Department of Veterans Affairs. *Med Care.* 1998;36(9):1324-36.

<sup>11</sup> Allison JJ, Wall TC, Spettell CM, Calhoun J, Fargason CA, Kobylinski RW, Farmer R, Kiefe C. The art and science of chart review. *The Joint Commission Journal on Quality Improvement.* 2000; 26(3):115-136.

### 3.3.4.2. Verifying Automated denominator case finding strategies (1A)

The Automated denominator case finding strategy ([Automated CFS](#)) should be applied to the performance measure patient population (Population) or a sample from the Population to identify cases to be included in the denominator. All patients in the Population or Population sample should be classified by the Automated CFS as belonging in the denominator (Automated CFS DEN+) or not belonging in the denominator (Automated CFS DEN-).

For the same Population or Population sample, a manual review of patient records should be conducted to identify cases for inclusion in the denominator. In the immediate context of this testing Protocol, manual review of patient medical records (whether in electronic or paper format) shall be assumed to be the reference strategy (gold standard) for case identification and performance measure calculation. All patients in the Population or Population sample should be classified by manual review as either belonging in the denominator ([Manual CFS DEN+](#)) or not belonging in the denominator (Manual CFS DEN-).

Cross-classification of patients in the sample by denominator inclusion status as determined by Automated case finding strategy and manual review yields the following 2x2 contingency table (Table 3.3.4.2):

**Table 3.3.4.2. Verifying EHR-based denominator case finding against reference strategy.**

|                    | Manual CFS DEN+ | Manual CFS DEN- |  |
|--------------------|-----------------|-----------------|--|
| Automated CFS DEN+ | TP              | FP              |  |
| Automated CFS DEN- | FN              | TN              | Total: Population (or population sample) |

- True positives (TP) are cases in the Population that were identified by Automated denominator case finding strategy (Automated CFS DEN+) for inclusion in the denominator AND were also identified by manual review denominator case finding strategy (Manual CFS DEN+) for inclusion in the denominator.
- False positives (FP) are cases that were identified by Automated CFS for inclusion in the denominator (Automated CFS DEN+) BUT were not identified by manual review for inclusion in the denominator (Manual CFS DEN-).
- True negatives (TN) are cases that were not identified by Automated CFS for inclusion in the denominator (Automated CFS DEN-) AND were not identified by manual review for inclusion in the denominator (Manual CFS DEN-).
- False negatives (FN) are cases that were not identified by Automated CFS for inclusion in the denominator (Manual CFS DEN-) BUT were identified by manual review for inclusion in the denominator (Manual CFS DEN+).

The sensitivity of an Automated case finding strategy can be calculated as:  $TP/(TP+FN)$ , and provides information on the hit rate of the Automated case finding strategy – i.e. the proportion of all cases that should have been included in the denominator that was identified by the Automated case finding strategy.

The specificity of an Automated case finding strategy can be calculated as:  $TN/(FP+TN)$ , and provides information on the ability of the strategy to discriminate – i.e. the proportion of all cases

that should not be included in the denominator that were not included on the basis of Automated case finding.

The accuracy of an Automated case finding strategy can be summarized by the following statistic:  $(TP+TN)/(TP+TN+FP+FN)$ . That is, accuracy of case finding strategy is the proportion cases that were true positives or true negatives. A perfectly accurate case finding strategy will yield only true positives or true negatives, and would yield an accuracy of 1:  $(TP+TN)/(TP+0+TN+0)$ .

The degree to which Automated CFS and Manual CFS yield concordant classifications may be summarized by Cohen's Kappa.

Confidence intervals may be constructed around estimates of the sensitivity, specificity, and kappa and formal statistical tests for significant differences in proportions identified by [Automated CFS](#) and [Manual CFS](#) should be conducted.<sup>13</sup> Reasons for discrepancies between Automated CFS and Manual CFS should be investigated and described qualitatively.<sup>14</sup>

Calculating the true negatives by reviewing every case in the population manually may present an undue burden to the researchers. This could be reduced by reviewing a proportion of the Automated CFS DEN+ and calculating a rate of true negatives. This rate could then be extrapolated to the population to estimate the number of true negatives to allow for calculation of specificity.

#### **3.3.4.3. Verifying numerator case finding within EHR- or Registry-based measurement (1B)**

For all cases in the patient sample that are identified by the Automated case finding strategy for inclusion in the denominator, apply the Automated numerator case-finding strategy to identify cases for inclusion in the numerator. All patients in the denominator should be classified as either belonging in the numerator (Automated CFS NUM+) or not belonging in the numerator (Automated CFS NUM-)

For all cases in the patient sample that are identified by the Manual case finding strategy for inclusion in the denominator, apply the Manual numerator case-finding strategy to identify cases to be included in the numerator. We will assume that manual review of patient records is the reference strategy (gold standard). All patients in the denominator should be classified by manual review as either belonging in the numerator (Manual CFS NUM+) or not belonging in the numerator (Manual CFS NUM-).

---

<sup>13</sup> For methodological examples, see: Kerr EA, Smith DM, Hogan MM, Krein SL, Pogach L, Hofer TP, Hayward RA. Comparing clinical automated, medical record, and hybrid data sources for diabetes quality measures. *The Joint Commission Journal on Quality Improvement*. 2002;28(10):555-565. Benin AL, Vitkauskas G, Thornquist E, Shapiro ED, Concato J, Aslan M, Krumholz HM. Validity of using an electronic medical record for assessing quality of care in an outpatient setting. *Medical Care*. 2005;43(7):691-698. Baker DW, Persell SD, Thompson JA, Soman NS, Burgner KM, Liss D, Kmetik KS. Automated review of electronic health records to assess quality of care for outpatients with heart failure. *Ann Intern Med*. 2007;146:270-277.

<sup>14</sup> For examples, see: Persell SD, Wright JM, Thompson JA, Kmetik KS, Baker DW. Quality measures for coronary artery disease using an electronic health record. *Arch Intern Med*. 2006;166:2272-2277. Baker DW, Persell SD, Thompson JA, Soman NS, Burgner KM, Liss D, Kmetik KS. Automated review of electronic health records to assess quality of care for outpatients with heart failure. *Ann Intern Med*. 2007;146:270-277.

Starting with a patient sample identified solely by manual review may not be practical in some projects. It is also an acceptable approach to start with all cases in the patient sample identified by the Automated case finding strategy for inclusion in the denominator, which were verified by manual review. This would remove patients who were found upon manual review to not meet the denominator inclusion criteria from the population sample evaluated for numerator inclusion.

Cross-classification of patients for inclusion in the denominator as determined by Automated case finding strategy and by manual review yields the following 2x2 contingency table (Table 3.3.4.3).

**Table 3.3.4.3. Verifying EHR-based numerator case finding vs. reference strategy**

|                    | Manual CFS NUM+ | Manual CFS NUM- |                        |
|--------------------|-----------------|-----------------|------------------------|
| Automated CFS NUM+ | TP              | FP              |                        |
| Automated CFS NUM- | FN              | TN              | Total: Manual CFS DEN+ |

- True positives (TP) are cases in the denominator that were identified by Automated CFS for inclusion in the numerator (Automated CFS NUM+) AND were also identified (i.e. confirmed) by manual review for inclusion in the numerator (Manual CFS NUM+).
- False positives (FP) are cases in the denominator that were identified by Automated CFS for inclusion in the numerator (Automated CFS NUM+) BUT were not identified by manual review for inclusion in the numerator (Manual CFS NUM-).
- True negatives (TN) are cases in the denominator that were not identified by Automated CFS for inclusion in the numerator (Automated CFS NUM-) AND were also not identified by manual review for inclusion in the numerator (Manual CFS NUM-).
- False negatives (FN) are cases in the denominator that were not identified by Automated CFS for inclusion in the numerator (Automated CFS NUM-) BUT were identified by manual review for inclusion in the numerator (Manual CFS NUM+).

The sensitivity of an Automated numerator case finding strategy can be calculated as:  $TP/(TP+FN)$ , and provides information on the hit rate of the Automated case finding strategy – i.e. the proportion of cases in the denominator that should have been included in the numerator, that were correctly identified by the Automated case finding strategy for inclusion in the numerator.

The specificity of an Automated numerator case finding strategy can be calculated as:  $TN/(FP+TN)$ , and provides information on the ability of the strategy to discriminate – i.e. the proportion of all cases in the denominator that should not be included in the numerator, that were correctly identified by the Automated numerator case finding strategy as ineligible for inclusion in the numerator.

The accuracy of an Automated numerator case finding strategy can be summarized by the following statistic:  $(TP+TN)/(TP+TN+FP+FN)$ . That is, accuracy of numerator case finding strategy is the proportion cases in the denominator that were correctly included in the numerator or correctly excluded from the numerator. A perfectly accurate case finding strategy will yield only true positives or true negatives, and would yield an accuracy of 1:  $(TP+TN)/(TP+0+TN+0)$ .

The degree to which Automated CFS and Manual CFS yield concordant classifications may be summarized by Cohen’s Kappa.

Confidence intervals may be constructed around estimates of the sensitivity, specificity, and kappa and formal statistical tests for significant differences in proportions identified by Automated CFS and Manual CFS should be conducted.<sup>15</sup> Reasons for discrepancies between Automated CFS and Manual CFS should be investigated and described qualitatively.<sup>16</sup>

#### **3.3.4.4. Verifying exception case finding within EHR- and Registry-based measurement (IC)**

For all cases in the patient sample that are identified by the Automated case finding strategy for inclusion in the denominator that were not identified by the Automated numerator case finding strategy for inclusion in the numerator - i.e. “apparent quality failures,” apply the Automated exception case-finding strategy to identify cases for denominator exception according to the exception criteria and measure specifications. All apparent quality failures should be classified by the EHR exception case finding strategy as either belonging in the excluded set (Automated CFS EXC+) or belonging in the denominator – i.e. not belonging in the excluded set (Automated CFS EXC-).

For all cases in the patient sample that are identified by the Manual case finding strategy for inclusion in the denominator that were not identified by the manual numerator case finding strategy for inclusion in the numerator - i.e. “apparent quality failures,” apply manual review of patient records to identify cases for denominator exception using the exception criteria and specifications for the measure in question. We will assume that manual review of patient records is the reference strategy (gold standard). All apparent quality failures should be classified by manual review as either belonging in the excluded set (Manual CFS EXC+) or belonging in the denominator – i.e. not belonging in the excluded set (Manual CFS EXC-).

Starting with a patient sample identified solely by manual review may not be practical in some projects. It is also an acceptable approach to start with all cases in the patient sample identified by the Automated case finding strategy for inclusion in the denominator but did not meet the numerator criteria, which were verified by manual review. This would remove patients who were manually found to not meet the denominator inclusion criteria or who did meet the numerator inclusion criteria from the population sample evaluated for exceptions. Cases found to meet the numerator using the automated numerator case finding strategy which were not verified by the manual numerator case finding strategy should be evaluated manually for exceptions.

---

<sup>15</sup> For methodological examples, see: Kerr EA, Smith DM, Hogan MM, Krein SL, Pogach L, Hofer TP, Hayward RA. Comparing clinical automated, medical record, and hybrid data sources for diabetes quality measures. *The Joint Commission Journal on Quality Improvement*. 2002;28(10):555-565. Benin AL, Vitkauskas G, Thornquist E, Shapiro ED, Concato J, Aslan M, Krumholz HM. Validity of using an electronic medical record for assessing quality of care in an outpatient setting. *Medical Care*. 2005;43(7):691-698. Baker DW, Persell SD, Thompson JA, Soman NS, Burgner KM, Liss D, Kmetik KS. Automated review of electronic health records to assess quality of care for outpatients with heart failure. *Ann Intern Med*. 2007;146:270-277.

<sup>16</sup> For examples, see: Persell SD, Wright JM, Thompson JA, Kmetik KS, Baker DW. Quality measures for coronary artery disease using an electronic health record. *Arch Intern Med*. 2006;166:2272-2277. Baker DW, Persell SD, Thompson JA, Soman NS, Burgner KM, Liss D, Kmetik KS. Automated review of electronic health records to assess quality of care for outpatients with heart failure. *Ann Intern Med*. 2007;146:270-277.

Cross classification of apparent quality failures for denominator exception as determined by Automated exception case finding strategy and manual review yields the following 2x2 contingency table (Table 3.3.4.4).

**Table 3.3.4.4. Verifying EHR-based exception case finding vs. reference strategy**

|                    | Manual CFS EXC+ | Manual CFS EXC- |  |
|--------------------|-----------------|-----------------|--|
| Automated CFS EXC+ | TP              | FP              |  |
| Automated CFS EXC- | FN              | TN              | Total: (Manual CFS DEN+) – (Manual CFS NUM+) |

- True positives (TP) are apparent quality failures that were identified by Automated CFS for denominator exception (Automated CFS EXC+) and that were also identified (i.e. confirmed) by manual review for denominator exception (Manual CFS EXC+).
- False positives (FP) are apparent quality failures that were identified by Automated CFS for denominator exception (Automated CFS EXC+) but that were not identified by manual review for inclusion in the numerator (Manual CFS EXC-).
- True negatives (TN) are apparent quality failures that were not identified by Automated CFS for denominator exception (Automated CFS EXC-) and that were also not identified by manual review for denominator exception (Manual CFS EXC-).
- False negatives (FN) are apparent quality failures that were not identified by Automated CFS for denominator exception (Automated CFS EXC-) but that were identified by manual review for denominator exception (Manual CFS EXC+).

The sensitivity of an Automated exception case finding strategy can be calculated as:  $TP/(TP+FN)$ , and provides information on the hit rate of the Automated exception case finding strategy – i.e. the proportion of all apparent quality failures should be excluded from the denominator that were correctly identified by the Automated exception case finding strategy for denominator exception.

The specificity of an Automated exception case finding strategy can be calculated as:  $TN/(FP+TN)$ , and provides information on the ability of the strategy to discriminate – i.e. the proportion of all apparent quality failures that should not be excluded from the denominator, that were correctly retained in the denominator on the basis of the Automated exception case finding strategy.

The accuracy of an Automated exception case finding strategy can be summarized by the following statistic:  $(TP+TN)/(TP+TN+FP+FN)$ . That is, accuracy of numerator case finding strategy is the proportion of apparent quality failures that were correctly classified by the Automated exception case finding strategy as eligible for denominator exception, and ineligible for denominator exception. A perfectly accurate exception case finding strategy will yield only true positives or true negatives, and would yield an accuracy of 1:  $(TP+TN)/(TP+0+TN+0)$ .

The degree to which Automated CFS and Manual CFS yield concordant classifications may be summarized by Cohen’s Kappa.

Confidence intervals may be constructed around estimates of the sensitivity, specificity, and Kappa and formal statistical tests for significant differences in proportions identified by

Automated CFS and Manual CFS should be conducted.<sup>17</sup> Reasons for discrepancies between Automated CFS and Manual CFS should be investigated and described qualitatively.<sup>18</sup>

### 3.3.4.5. Verifying overall performance using Automated strategies vs. reference strategy (1D)

In addition to verifying componential Automated case finding strategies, overall calculations of PCPI measures based on Automated case-finding strategies should be compared to overall calculations of PCPI measures based on manual review of medical records. The purpose of this test is to assess the distortion in overall performance calculations due to inaccuracies in Automated case finding strategies. Table 3.3.4.5 provides an overview of the calculations and comparisons involved in verifying Automated performance measures vs. measures derived from a reference strategy (manual review of patient records).

**Table 3.3.4.5. Verifying overall EHR-based performance measures vs. reference strategy**

|                        | Automated Measurement   | Complete Manual Review  | Difference                 |
|------------------------|---|---|----------------------------|
| <b>DENOMINATOR</b>     | Apply Automated CFS (denominator) to SAMPLE to identify denominator<br>$DEN_{AUTO}$   | Apply Manual CFS (denominator) to SAMPLE to identify denominator<br>$DEN_{MAN}$                 | $DEN_{AUTO} - DEN_{MAN}$   |
| <b>NUMERATOR</b>       | Apply Automated CFS (numerator) to $DEN_{AUTO}$ to identify numerator<br>$NUM_{AUTO}$ | Apply Manual CFS (numerator) to $DEN_{MAN}$ to identify numerator<br>$NUM_{MAN}$                | $NUM_{AUTO} - NUM_{MAN}$   |
| <b>EXCEPTION</b>       | Apply Automated exception CFS to $DEN_{AUTO}$ to identify numerator<br>$EXC_{AUTO}$   | Apply Manual CFS (exception) to $(DEN_{MAN} - NUM_{MAN})$ to identify exceptions<br>$EXC_{MAN}$ | $EXC_{AUTO} - EXC_{MAN}$   |
| <b>OVERALL Measure</b> | $PERF_{AUTO} = NUM_{AUTO} / (DEN_{AUTO} - EXC_{AUTO})$                                | $PERF_{MAN} = NUM_{MAN} / (DEN_{MAN} - EXC_{MAN})$  | $PERF_{AUTO} - PERF_{MAN}$ |

<sup>17</sup> For methodological examples, see: Kerr EA, Smith DM, Hogan MM, Krein SL, Pogach L, Hofer TP, Hayward RA. Comparing clinical automated, medical record, and hybrid data sources for diabetes quality measures. *The Joint Commission Journal on Quality Improvement*. 2002;28(10):555-565. Benin AL, Vitkauskas G, Thornquist E, Shapiro ED, Concato J, Aslan M, Krumholz HM. Validity of using an electronic medical record for assessing quality of care in an outpatient setting. *Medical Care*. 2005;43(7):691-698. Baker DW, Persell SD, Thompson JA, Soman NS, Burgner KM, Liss D, Kmetik KS. Automated review of electronic health records to assess quality of care for outpatients with heart failure. *Ann Intern Med*. 2007;146:270-277.

<sup>18</sup> For examples, see: Persell SD, Wright JM, Thompson JA, Kmetik KS, Baker DW. Quality measures for coronary artery disease using an electronic health record. *Arch Intern Med*. 2006;166:2272-2277. Baker DW, Persell SD, Thompson JA, Soman NS, Burgner KM, Liss D, Kmetik KS. Automated review of electronic health records to assess quality of care for outpatients with heart failure. *Ann Intern Med*. 2007;146:270-277.

Performance measures should be calculated using denominators, numerators, and exceptions identified through automated case finding strategies and algorithms as applied to the Population or Population sample. Independently, performance measures should be calculated using denominators, numerators, and exceptions identified through manual review as applied to all cases in the Population or Population sample. The Automated CFS-derived denominators, numerators and exceptions ( $DEN_{AUTO}$ ,  $NUM_{AUTO}$ ,  $EXC_{AUTO}$ ) are the same as those identified in the componential tests 1A through 1C as described previously in Sections 3.3.1.2-3.3.1.4. However, *the denominators, numerators, and exceptions to be identified by manual review ( $DEN_{MAN}$ ,  $NUM_{MAN}$ ,  $EXC_{MAN}$ ) for calculating an overall gold standard performance measure should be calculated independently when practical, starting from the [Population](#) or [Population sample](#).*

Formal statistical tests for differences in proportions should be performed. Reasons for discrepancies between Automated components/measures and components/measures derived from manual review should be investigated and documented.

#### **3.3.4.6. Process Evaluation**

In addition to reporting the results of denominator, numerator, and exception verification as discussed in Sections 3.3.4.1-3.3.4.5, investigators should provide additional evaluation of their EHR- and registry-based measurement strategy and process. At a minimum, a process evaluation of EHR- and registry-based measurement should include the following:

- Thorough description of the process by which PCPI measure definitions and specifications were integrated into the EHR or registry, if applicable
- Description of the following processes as well as process flow:
  - Identification of Population
    - Where appropriate: sampling procedure for extracting Population sample
  - Denominator case finding
  - Numerator case finding
  - Exception case finding
  - Measure performance rate and exception rate calculation
  - Measure performance rate and exception rate reporting
  - Data management/export (where applicable)
- Description of manual review process
  - Selection of reviewer(s) and reviewer qualifications and experience
  - Reviewer training
    - Trainer
    - Duration of training
    - Assessment of competence implementing the paper medical record based measurement strategy
    - Inter-abstractor reliability achieved after training, prior to formal review of records
  - Abstraction procedures
    - Procedures for adjudication between reviewer discrepancies
- Evaluation of the quality of individual data elements retrieved through automated procedures compared to manual review. Forms that may be used to facilitate the reporting of the quality of individual data elements may be found in the exhibits at the end of this section. These forms provide a structured means for reporting: discrepancies in

individual data elements as found through automated data retrieval/abstraction vs. manual review; missing data; barriers encountered in retrieving data elements through automated processes.

### **3.3.5. Methodology (Parallel Forms Reliability): Automated Measurement Modalities vs. Manual Review**

#### **3.3.5.1. Verifying administrative/claims data-based denominator case finding strategies (2A)**

The automated administrative/claims data-based case finding strategy (Automated CFS) should be applied to the performance measure patient population ([Population](#)) or a sample from the Population to identify cases to be included in the denominator. All patients in the Population or Population sample should be classified by the Administrative CFS as belonging in the denominator (Automated CFS DEN+) or not belonging in the denominator (Automated CFS DEN-).

For the same Population or Population sample, a manual review of patient records should be conducted to identify cases for inclusion in the denominator. In the immediate context of this testing Protocol, manual review of patient medical records (whether in electronic or paper format) shall be assumed to be the reference strategy (gold standard) for case identification and performance measure calculation. All patients in the Population should be classified by manual review as either belonging in the denominator (Automated CFS DEN+) or not belonging in the denominator (Automated CFS DEN-).

The same methodologies used for Electronic Health Records and Registry-based parallel forms reliability can be utilized for Administrative/claims data-based parallel forms reliability. Please refer to section 3.3.4 for a thorough explanation of these strategies.

#### **3.3.5.2. Process Evaluation**

In addition to reporting the results of denominator, numerator, and exception verification as discussed in Sections 3.3.4.2-3.3.4.5, investigators should provide additional evaluation of their administrative/claims-based measurement strategy and process. At a minimum, a process evaluation of administrative/claims-based measurement should include the following:

- Description of the following processes as well as process flow;
  - Identification of Population
    - Where appropriate: sampling procedure for extracting Population sample
  - Denominator case finding
  - Numerator case finding
  - Exception case finding
  - Measure performance rate and exception rate calculation
  - Measure performance rate and exception rate reporting
  - Data management/export (where applicable)
- Description of manual review process;
  - Selection of reviewer(s) and reviewer qualifications and experience
  - Reviewer training
    - Trainer

- Duration of training
  - Assessment of competence implementing the paper medical record based measurement strategy
  - Inter-abstractor reliability achieved after training, prior to formal review of records
- And abstraction procedures
  - Procedures for adjudication between reviewer discrepancies
- Evaluation of the quality of individual data elements retrieved through the administrative/claims-based procedures compared to manual review. Forms that may be used to facilitate the reporting of the quality of individual data elements may be found in the exhibits at the end of this section. These forms provide a structured means for reporting: discrepancies in individual data elements as found through administrative/claims-based strategies for data retrieval/abstraction vs. manual review; missing data; and barriers encountered in retrieving data elements through administrative/claims-based processes.

### 3.4. Other Forms of Reliability Testing

#### 3.4.1. Test-retest reliability

##### 3.4.1.1. Definition

*Test-retest reliability* refers to the extent to which a survey or measurement instrument elicit the same response from the same respondent across short intervals of time: the stability of a survey or measurement instrument’s “performance.” To assess test-retest reliability, the same instrument is applied to the same observations across short intervals of time.

##### 3.4.1.2. Recommendation

The PCPI MIE does not recommend assessments of test-retest reliability for PCPI performance measures.

**Recommendation 3-4. Assessments of test-retest reliability for PCPI performance measures are not recommended.**

##### 3.4.1.3. Rationale

Test-retest reliability is generally not applicable to the evaluation of PCPI performance measures. Because data on physician behavior take the form of documented observations in EHR or paper charts and/or administrative/claims databases, physician performance in treating a specific patient cannot be reproduced the way a respondent can answer the same test questions twice. Because available data on physician performance over a measurement period are immutable, historical facts, repeating data collection and performance measurement procedures on static data over small intervals of time – particularly when data collection and measurement procedures are automated or highly standardized – is unlikely to yield valuable information that would justify the cost and burden of replication.

### 3.4.2. *Internal Consistency*

#### 3.4.2.1. *Definition*

A multiple-item test or survey instrument exhibits *internal consistency* to the extent that the items designed to measure a given construct are inter-correlated.

#### 3.4.1.2. *Recommendation*

The PCPI MIE does not recommend assessments of internal consistency for non-composite PCPI performance measures.

**Recommendation 3-5. Assessments of internal consistency for non-composite PCPI performance measures are not recommended.**

#### 3.4.2.3. *Rationale*

PCPI measure sets are comprised of two or more individual physician-level measures that have been identified by a panel of clinical and technical experts as indicative of quality of care within a given clinical (topical) domain. Evaluation of internal consistency is not applicable to the testing of individual PCPI measures. In theory, if individual measures within a measure set constitute valid components of a single concept of what is to be regarded as quality of healthcare within a given clinical domain, then we should expect physician performance on one measure to predict performance on another measure within the same set. In practice, however, physician performance is not measured at the level of the measure set – there are currently no developed and ratified composite measures aggregating performance across multiple indicators into a summary index. Moreover, physicians reporting on one indicator within a measure set may not report on all measures within the measure set. Currently, measure sets function as collections of individual, discrete performance measures within a broad topical domain. As such, the focus of measure testing is on the scientific soundness of individual measures, and tests of internal consistency are not required. However, in the event that composite measures are developed that aggregate two or more individual PCPI measures, then tests of internal consistency may be recommended.

## 3.5. **Scope of Reliability Testing**

### 3.5.1. *Recommendation*

The PCPI MIE recommends that at a minimum, the reliability of each PCPI measure should be formally evaluated in: (a) private practice settings; (b) publicly funded (federal/state/county/municipal) settings that serve disproportionate shares of clinically and/or sociodemographically vulnerable patients; (c) practices that use EHR-based performance measurement modalities; (d) practices that use administrative/claims data-based measurement modalities; and (e) practices that use paper medical record data-based measurement modalities.

***Recommendation 3-6. When feasible, it is strongly recommended that the reliability of each PCPI measure be formally evaluated in: (a) private practice settings; (b) publicly funded (federal/state/county/municipal) settings that serve disproportionate shares of clinically and/or sociodemographically vulnerable patients; (c) practices that use EHR-based performance measurement modalities; (d) practices that use administrative/claims data-based measurement modalities; and (e) practices that use paper medical record data-based measurement modalities.***

### ***3.5.2. Remarks***

It is recognized that there is great heterogeneity in EHR products, and even greater heterogeneity is to be expected in the integration and implementation of PCPI measures within EHR products. Mandating demonstration of satisfactory metrical performance for each component of each measure in all EHR products is not possible. There is a possibility that a measure may perform satisfactorily in one EHR-based study, and perform less well on another EHR-based study using a different EHR product and/or a different method of measure integration and implementation. In such circumstances, it is the recommendation of the PCPI Measures Evaluation and Implementation Advisory Committee that, if one study demonstrates a measure to perform satisfactorily only to be contradicted by a subsequent study using a different EHR product and/or implementation method, the measures shall be deemed metrically satisfactory (on the basis of the first study) until an exact replication of the first study's methods yields findings that demonstrate the measure to be unsatisfactory.

**Exhibit 6. Example Concordance Table for Data Elements<sup>19</sup>**

| Data Element   | Reabstraction Sample | Total Discrepancies | Concordance (% Agreement) | Reason for Discrepancies |   |   |   |   |   |   |   |
|----------------|----------------------|---------------------|---------------------------|--------------------------|---|---|---|---|---|---|---|
|                |                      |                     |                           | 1                        | 2 | 3 | 4 | 5 | 6 | 7 |   |
| Data Element 1 |                      |                     |                           |                          |   |   |   |   |   |   |   |
| Data Element 2 |                      |                     |                           |                          |   |   |   |   |   |   |   |
| •              | •                    | •                   | •                         | •                        | • | • | • | • | • | • | • |
| •              | •                    | •                   | •                         | •                        | • | • | • | • | • | • | • |
| •              | •                    | •                   | •                         | •                        | • | • | • | • | • | • | • |
| Data Element D |                      |                     |                           |                          |   |   |   |   |   |   |   |

Reason for Discrepancies: 1=Data entry/transcription error; 2=Information missed; 3=Illegible document; 4=Conflicting information; 5=Unclear element definition; 6=Not following definition; 7=Other/not determined.

**Exhibit 7. Example Table for Kappa Statistics with Confidence Intervals**

| MEASURE/<br>MEASURE TYPE                          | Y/Y | Y/N | N/Y | N/N | Kappa | 95% CI |
|---|-----|-----|-----|-----|-------|--------|
| <i>Measure Set ( __ patient records reviewed)</i> |     |     |     |     |       |        |
| <i>Measure 1 denominator</i>                      | X   | X   | X   | X   | X     | X      |
| <i>Measure 1 numerator</i>                        | X   | X   | X   | X   | X     | X      |
| <i>Measure 1 exceptions</i>                       | X   | X   | X   | X   | X     | X      |
| <i>Measure 1 overall*</i>                         | X   | X   | X   | X   | X     | X      |
| <i>Measure 2 denominator</i>                      | X   | X   | X   | X   | X     | X      |
| •   | •   | •   | •   | •   | •     | •      |
| •   | •   | •   | •   | •   | •     | •      |
| •   | •   | •   | •   | •   | •     | •      |
| <i>Measure n overall</i>                          | X   | X   | X   | X   | X     | X      |

In the above table, the Y/Y indicates that both abstractors answered “yes”; Y/N indicates that abstractor 1 answered “yes” and abstractor 2 answered “no”; N/Y indicates that abstractor 1 answered “no” and abstractor 2 answered “yes”; and N/N indicates that both abstractors answered “no”.

\* The overall kappa refers to the kappa score calculated if the gold standard and the re-abstraction results are compared as follows:

1 = measure met

2 = measure not met, no exception

3 = exception, measure not met

4 = not eligible for denominator

| Patient | Gold Standard Abstraction | Abstraction of Alternative Data Modality |
|---------|---------------------------|--|
| A       | 1                         | 1  |
| B       | 1                         | 2  |
| C       | 2                         | 2  |
| D       | 1                         | 3  |

<sup>19</sup> Modeled after tables in: Doctor’s Office Quality Final Report. Project 2002-2005 Coordinating QIO. Iowa Foundation for Medical Care. December 2005.

**Exhibit 8. Example Diagnostic Audit Table for Discrepant Data Elements<sup>20</sup>**

| <b>Data Element</b> | <b>Reason for Discrepancy</b>     | <b>Freq.</b> | <b>Pct. (%)</b> | <b>Barriers/ Description</b> |
|---------------------|-----------------------------------|--------------|-----------------|------------------------------|
| Data Element 1      | 1. Data entry/transcription error |              |                 |                              |
|                     | 2. Information missed             |              |                 |                              |
|                     | 3. Illegible document             |              |                 |                              |
|                     | 4. Conflicting information        |              |                 |                              |
|                     | 5. Unclear element definition     |              |                 |                              |
|                     | 6. Not following definition       |              |                 |                              |
|                     | 7. Other/Not determined           |              |                 |                              |
|                     | TOTAL                             |              |                 |                              |
| Data Element 2      | 1. Data entry/transcription error |              |                 |                              |
|                     | 2. Information missed             |              |                 |                              |
|                     | 3. Illegible document             |              |                 |                              |
|                     | 4. Conflicting information        |              |                 |                              |
|                     | 5. Unclear element definition     |              |                 |                              |
|                     | 6. Not following definition       |              |                 |                              |
|                     | 7. Other/Not determined           |              |                 |                              |
|                     | TOTAL                             |              |                 |                              |
| •                   | •                                 | •            | •               | •                            |
| •                   | •                                 | •            | •               | •                            |
| •                   | •                                 | •            | •               | •                            |
| Data Element D      | 1. Data entry/transcription error |              |                 |                              |
|                     | 2. Information missed             |              |                 |                              |
|                     | 3. Illegible document             |              |                 |                              |
|                     | 4. Conflicting information        |              |                 |                              |
|                     | 5. Unclear element definition     |              |                 |                              |
|                     | 6. Not following definition       |              |                 |                              |
|                     | 7. Other/Not determined           |              |                 |                              |
|                     | TOTAL                             |              |                 |                              |

<sup>20</sup> Modeled after tables in: Doctor's Office Quality Final Report. Project 2002-2005 Coordinating QIO. Iowa Foundation for Medical Care. December 2005.

## SECTION IV. TESTING AREA 4: VALIDITY

### Section Outline

- 4.1. [Validity and PCPI Performance Measures](#)
  - 4.1.1. [Definition](#)
- 4.2. [Face Validity](#)
  - 4.2.1. [Definition](#)
  - 4.2.2. [Recommendation](#)
  - 4.2.3. [Rationale](#)
- 4.3. [Content Validity](#)
  - 4.3.1. [Definition](#)
  - 4.3.2. [Recommendation](#)
  - 4.3.3. [Rationale](#)
- 4.4. [Construct Validity](#)
  - 4.4.1. [Definition](#)
  - 4.4.2. [Recommendation](#)
  - 4.4.3. [Rationale](#)

### **Priority II Testing**

- 4.5. [Predictive Validity](#)
  - 4.5.1. [Definition](#)
  - 4.5.2. [Recommendation](#)
  - 4.5.3. [Rationale](#)
  - 4.5.4. [Methodology](#)
- 4.6. [Scope of Predictive Validity Testing](#)
  - 4.6.1. [Recommendation](#)

## **4.1. Validity and PCPI Performance Measures**

### **4.1.1. Definition**

The validity of a PCPI performance measure refers to the extent to which it truly measures that which it is intended and designed to measure. Four common types of validity that are relevant to the evaluation of PCPI measures are face, content, construct, and predictive validity.

## **4.2. Face Validity**

### **4.2.1. Definition**

*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 a measure set poses the question of how well the individual indicators comprising the measure set seem to reflect quality of care within the given clinical topical domain. For example, in the case of the PCPI/ACC/AHA Heart Failure Measure Set, do constituent individual indicators taken together appear to reflect what is to be considered quality of care in Heart Failure?

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 healthcare quality as intended. In the case of the individual measure, ACE Inhibitor or ARB Therapy in Heart Failure, face validity refers to whether numerator, denominator, and exception definitions appear to capture the rate at which a physician appropriately prescribes to ACE/ARB therapy to heart failure patients in a manner that is indicative of quality of care.

### **4.2.2. Recommendation**

The PCPI MIE recommends that all PCPI performance measures derived from evidence-based guidelines can be assumed to have face validity: no additional testing to establish face validity is required. For measures based on expert opinion, or measures which will have a wide scope of implementation not fully represented by the measure Work Group, the measure Work Group may request that additional face validity testing be conducted.

***Recommendation 4-1. All PCPI performance measures are assessed for face validity by expert Work Group members during the development process. No additional testing to establish face validity is required, unless requested by the workgroup.***

### **4.2.3. Rationale**

It is the consensus of the PCPI MIE that face validity of PCPI measures can be assumed to be established once they have progressed beyond the Public Comment period by virtue of the specialized expertise of the PCPI Measure Development Work Group members who are involved in identifying and drafting performance measures within a topical domain as well, as the rigorous, structured discussions that are prescribed according to PCPI protocols for Work Group conduct.

#### **4.2.4. Pilot testing for face validity**

For measures based on expert opinion, one option is to perform pilot testing of the measures with focus groups or using questionnaires. Sample questions for this type of testing effort might include the following, rated on a scale of 1-5, from strongly disagree to strongly agree.

1. Do you believe that there is sufficient evidence to support the use of this measure at a widespread level?
2. Does a significant opportunity for improvement exist; measure addresses an area(s) where there is a substantial gap between optimal and current clinical practice?
3. Are the measure rationale and results easily understood by and meaningful to users of the data?
4. How often would information derived from this measure raise good questions for use in your hospital's quality improvement activities?
5. Is improvement under provider control; does health care organization have the responsibility, control, and ability to effect change of the processes and/or outcomes being measured?

### **4.3. Content validity**

#### **4.3.1. Definition**

*Content validity* is the “extent to which an empirical measurement reflects a specific domain of content.”<sup>21</sup>

At the level of the measure set, content validity pertains to the extent to which the individual indicators collectively reflect elements of care that are germane to the overall quality of care within a given clinical topical domain. For example, the content validity of the PCPI/ACC/AHA Heart Failure Measure Set refers to the extent to which the following services all reflect what is considered to be quality in caring for patients with Heart Failure: left ventricular function assessment, weight measurement, blood pressure measurement, volume overload clinical symptoms assessment, patient education, beta-blocker therapy, ACE/ARB therapy, warfarin therapy for patients with atrial fibrillation, and laboratory testing

At the level of an individual measure, content validity pertains to the extent to which the measure definition and its specifications truly capture and reflect a given element of care (in the case of process measures), structural or organizational feature (in the case of structural measures), or outcome (in the case of outcome measures). For example, the content validity of the Laboratory Tests measure in the PCPI Heart Failure Measure Set refers to the extent to which the measure definition and specification adequately capture all aspects of laboratory testing that are characteristic of appropriate, high-quality care for heart failure patients (on the basis of clinical evidence). In this example, a laboratory tests measure defined only on the basis of performing a complete blood count would have lower content validity than a measure defined on the basis of performing a complete blood count, urinalysis, serum electrolytes, serum creatinine, blood urea nitrogen, liver function tests, and thyroid-stimulating hormone.<sup>22</sup>

#### **4.3.2. Recommendation**

---

<sup>21</sup> Carmines EG and Zellner RA. *Reliability and Validity Assessment*. London: Sage, 1979. Page 20.

<sup>22</sup> See <http://www.ama-assn.org/ama1/pub/upload/mm/370/hfset-12-5.pdf>

The PCPI MIE recommends that all PCPI performance measures derived from evidence-based guidelines can be assumed to have content validity: no additional testing to establish content validity is required. For measures based on expert opinion, or measures which will have a wide scope of implementation not fully represented by the measure workgroup, the measure workgroup may request that additional content validity testing be conducted.

**Recommendation 4-2. All PCPI performance measures are assessed for content validity by expert Work Group members during the development process. No additional testing to establish content validity is required, unless requested by the workgroup.**

#### 4.3.3. Rationale

It is the consensus of the PCPI MIE that content validity of PCPI measures can be assumed to be established once they have progressed beyond the Public Comment period by virtue of the specialized expertise of the PCPI Work Group members who are involved in identifying and drafting performance measures within a topical domain as well, as the rigorous, structured discussions that are prescribed according to PCPI protocols for Work Group conduct.

### 4.4. Construct validity

#### 4.4.1. Definition

In common psychometric context, *construct validity* typically refers to the extent to which the operationalization of a measure reflects the underlying abstract concept that it is designed to reflect. There are two subcategories of construct validity:

- *Convergent validity.* Convergent validity refers to the degree to which multiple indicators of a single underlying concept are correlated. Convergent validity is typically a property of a multiple-item test, or a multiple-indicator composite measure.
- *Discriminant validity.* Given two distinct constructs, discriminant validity refers to the degree to which there are no cross-associations between constituents of one construct with those of the other. Cross-associations suggest that the operationalization of a construct does not measure only that which it was designed to measure, and may thus indicate the need to further refine the definition and/or specification of the constructs.

Convergent validity is primarily a property of measurement sets rather than individual measures. For example, to the extent that individual measures comprising the PCPI/ACC/AHA CAD measure set all reflect quality of care for CAD rather than quality of care for Community Acquired Pneumonia (CAP), CAD measures should not be strongly associated with CAP measures.

Discriminant validity will primarily be of interest for measures in the same measurement set but also may be of interest in similar measures across measurement sets. Information gained from testing discriminant validity, while valuable, is not considered necessary at this stage of measure maturity.

#### **4.4.2. Recommendation**

The PCPI MIE recommends that all PCPI performance measures be exempt from construct validity testing.

**Recommendation 4-3. All individual PCPI performance measures are exempt from construct validity testing.**

#### **4.4.3. Rationale**

Convergent and discriminant validity are more appropriately considered properties of *measure sets* rather than *individual* measures. Because current specifications of PCPI measures do not provide methodologies for the calculation of composite scores at the level of measure sets, the PCPI MIE does not consider construct validity testing to be of immediate relevance and recommends that PCPI measures be exempt from requirements requesting documentation of construct validity testing.

## **PRIORITY II TESTING**

## 4.5. Predictive validity

### 4.5.1. Definition

The *predictive validity* of a test or measure refers to the presence of association between the given test (or measure) and relevant outcomes of interest.

In the context of PCPI performance measurement, a measure with predictive validity would be expected to be strongly associated with patient clinical outcomes (or intermediate outcomes) and/or patient satisfaction. For example, an evaluation of the predictive validity of the Timing of Prophylactic Antibiotics (Ordering Physician) from the PCPI Perioperative Care Measure Set might assess:

- At the patient level, whether receipt of prophylactic antibiotics within one hour of surgery predicts better survival, shorter length of stay, and/or reduced risk of complications, readmission, or other appropriate indicators of clinical outcomes
- At the physician level, whether physician performance (adherence) rates predict rates of patient mortality, average length of stay, and/or rates of relevant and appropriate patient outcomes.

### 4.5.2. Recommendation

For individual measures and measure sets focusing on processes of care that are supported by a strong, clinical evidence base, studies of predictive validity are of value but dependant upon data availability and the stage of implementation. For individual measures and measure sets focusing on processes of care that are not supported by a clear, conclusive evidence base, studies of predictive validity are recommended where possible.

***Recommendation 4-4. Predictive validity testing is recommended where possible for all PCPI measures focusing on processes of care that are not supported by a strong evidence base. Predictive validity testing is of value but dependant upon data availability and the stage of implementation for all PCPI measures focusing on processes of care that are founded upon a strong, conclusive evidence base.***

### 4.5.3. Rationale

It has also been argued that studies of measure predictive validity may result in unnecessary use of resources to pursue research that would be duplicative of the randomized, controlled trials or other clinical studies from which guidelines are derived, and in turn upon which many PCPI measures are based. Insofar as it is the intent of PCPI that all PCPI measures be evidence-based (derived from clinical practice guidelines) the existence of a strong, valid association between the process, structure, or outcome of care upon which a measure is based can be assumed to hold. The diversion of additional resources as well as the potential delay in measure approval that may ensue from the initiation of additional studies to reestablish the link between process measures and relevant outcomes may be unwarranted. However, in cases where the strength of the evidence underlying a PCPI measure may be uncertain, then studies of predictive validity for that

measure will be compulsory. The decision as to whether a measure must be tested for predictive validity will be made on a measure-by-measure basis by the PCPI and the relevant Measure Development Work Group.

In cases where measures are founded upon uncertain evidence or consensus-based guidelines, tests of predictive validity are of value. The MIE encourages studies of predictive validity for two primary reasons. First, while information derived from process-of-care and structural measures may particularly useful for monitoring and improving quality, the utility of such information may be less transparent for other stakeholders such as healthcare consumer,<sup>23</sup> and to a lesser degree, purchasers and payers. Decisions of the consumer public, and possibly purchasers and payers as well, may be more responsive to information regarding performance and its effect on outcomes, relative to information on performance levels alone. For this reason, studies assessing the predictive validity of PCPI measures are highly desirable.

Second, it must be remembered that randomized controlled trials generally demonstrate the efficacy of certain treatments or processes, not effectiveness. Because population performance measurement will be applied in uncontrolled settings and circumstances, it would be valuable for the sake of measure development and performance management to know whether and to what degree a performance measure can predict outcomes or other events of interest. To date, there have been only a handful of studies testing the association between some form of performance measure and outcomes, and these have generally been studies of inpatient processes at the hospital level. These studies have found very small, if any associations, between process measures and patient outcomes.<sup>24,25</sup> This may be related to differences in implementation processes (or treatments) and undiscovered confounders in effectiveness studies compared with efficacy studies.

#### ***4.5.4. Methodology: Predictive Validity***

##### ***4.5.4.1. Dimensions of Patient Outcomes***

Relevant patient outcomes to assess may fall under 3 broad categories::

1. Clinical/physiological changes in disease (or health) status
2. Health-related quality of life
3. Patient satisfaction

Clinical and or physiological changes in disease or health status can be marked by two types of measures: intermediate outcome measures as well as true outcome measures. Intermediate outcome measures are clinically observable and measurable markers of the symptoms and or physical factors inherent to the specific condition being measured. An example of an intermediate outcome measure is the reduction in HbA1c values among patients with diabetes. A

---

<sup>23</sup> Rubin HR, Pronovost P, Diette GB. The advantages and disadvantages of process-based measure of health care quality. *International Journal for Quality in Health Care*. 2001;13(6):469-474.

<sup>24</sup> Bradley EH, Herrin J, Elbel B, McNamara RL, Magid DJ, Nallamothu BK, Wang Y, Normand ST, Spertus J, Krumholz HM. Hospital quality for acute myocardial infarction: correlation among process measures and relationship with short-term mortality. *JAMA*. 2006;296:72-78.

<sup>25</sup> Werner RM, Bradlow ET. Relationship between Medicare's Hospital Compare performance measures and mortality rates. *JAMA*. 2006;296:2694-2702.

true outcome measure assesses the result of the patients' episode of care. An example of an outcome measure is the 30-day readmission rate for heart failure patients.

Health-related quality of life is an approach to defining patient outcomes that focus on a patient's subjective evaluation of his/her physical and mental health and physical and social functional ability. An example of patient health-related quality of life assessment tool is the RAND Medical Outcomes Study Short Form 36 instrument<sup>26</sup>.

Patient satisfaction indicators capture the subjective evaluation of patients with respect to their healthcare experience. An example of a patient satisfaction outcome assessment tool is the Agency for Healthcare Research and Quality Consumer Assessment of Healthcare Providers and Systems (CAHPS) survey<sup>27</sup>.

We recommend that outcomes research involving PCPI measures cover measures that are meaningful for patients and physicians, as well as other stakeholders (e.g. planners and payers).

#### ***4.5.4.2. Types of Outcome Measures***

Two broad categories of outcome measures exist: condition-specific outcome measures, and global (or 'generic') outcome measures.

1. *Condition-specific outcome measures.* Condition-specific outcome measures are measures of patient physical, mental, and/or functional well-being that are designed to reflect the direct results of treatment for a specific health problem or condition.
2. *Global outcome measures.* Global (or generic) outcome measures are measures of a patient's overall patient physical, mental, and/or functional well-being that are not designed. Global outcome measures can be used across different clinical and topical domains, and may facilitate comparison across individual indicators.

Condition-specific measures of outcome may be more sensitive in reflecting differences in clinical performance and information derived from them may be of particular value to clinicians and healthcare professionals in the context of quality improvement cycles; however, these measures may be less transparent and meaningful to consumers and other stakeholders. More research is needed in assessing the sensitivity of global measures to variations in process measures. In consideration of time and resource constraints in measure testing, the MIE recommends that, when tests of predictive validity are conducted, condition-specific measures of patient outcomes are used.

#### ***4.5.4.3. Identification of Condition-Specific Outcome Measures***

The MIE recommends that PCPI Measure Development Work Groups (hereafter, Work Groups) be responsible for identifying a short list of outcome measures for use with each performance measure set. As discussed above, this short list should include:

- Measures of clinical outcomes, functional outcomes, and patient satisfaction

---

<sup>26</sup>The RAND Medical Outcomes Study Short Form 36 instrument, available at: [http://www.rand.org/health/surveys\\_tools/mos/mos\\_core\\_36item.html](http://www.rand.org/health/surveys_tools/mos/mos_core_36item.html)

<sup>27</sup> The Agency for Healthcare Research and Quality Consumer Assessment of Healthcare Providers and Systems (CAHPS) survey, available at <https://www.cahps.ahrq.gov/default.asp>

- Condition-specific outcomes and global outcomes

The identification of standard, evidence-based outcomes should be based on the most recent, relevant literature available. In selecting outcome measures, consideration should be directed to: the strength and quality of evidence linking similar process-of-care and/or structural measures to outcomes; the feasibility of data collection for outcomes measurement; and the sensitivity of the outcomes measure with respect to changes or differences in process-of-care and/or structure.

The short list of outcomes measures may be instituted within the Work Group measure development process and undergo periodic review and revision when PCPI measures are reviewed and revised every four years. Reasons for subsuming this task within the Work Group measure development process include the following:

- The identification of appropriate, standard, evidence-based outcomes measures requires the specialized substantive and methodological expertise of Work Groups.
- The discussion of outcomes is germane to establishing the clinical importance and construct validity of performance measures, and thus falls within the content purview of Work Groups.
  - Embedding the identification of outcomes measures to be used in testing performance measures within the Work Group measure development process will make it unnecessary to convene additional PCPI task groups with redundant expertise and increase efficiency in overall development and testing.

Identifying appropriate, standard, evidence-based outcomes measures at the time of measure development will expedite measure testing by enabling studies of predictive

## **4.6. Scope of Predictive Validity Testing**

### ***4.6.1. Recommendation***

It is desirable that assessments of predictive validity (where indicated) be carried out in diverse settings, including: (a) private practice settings; (b) publicly funded (federal/state/county/municipal) settings that serve disproportionate shares of clinically and/or sociodemographically vulnerable patients; (c) practices that use EHR-based performance measurement modalities; (d) practices that use administrative/claims data-based measurement modalities; and (e) practices that use paper medical record data-based measurement modalities

***Recommendation 4-5. It is desirable for PCPI measures to demonstrate predictive validity studies in diverse settings, including: (a) private practice settings; (b) publicly funded (federal/state/county/municipal) settings that serve disproportionate shares of clinically and/or sociodemographically vulnerable patients; (c) practices that use EHR-based performance measurement modalities; (d) practices that use administrative/claims data-based measurement modalities; and (e) practices that use paper medical record data-based measurement modalities.***

## SECTION V. TESTING AREA 5: UNINTENDED CONSEQUENCES

### Section V Outline

- 5.1. [Definition](#)
- 5.2. [Types of Unintended Consequences Arising from the Use of Performance Measures in Reporting and/or Incentive Programs](#)
  - 5.2.1. [Unintended Consequences at the Physician/Physician-Patient Level](#)
  - 5.2.2. [Unintended Consequences at the Organizational Level](#)
  - 5.2.3. [Unintended Consequences at the Aggregate Level](#)
  - 5.2.4. [Positive unintended consequences](#)
- 5.3. [Factors Influencing the Occurrence of Unintended Consequences](#)
- 5.4. [Unintended Consequences and PCPI Performance Measures](#)
  - 5.4.1. [Recommendation](#)
  - 5.4.2. [Rationale](#)
- 5.5. [General Considerations for Investigating Unintended Consequences](#)
  - 5.5.1. [Patient Selection at the Extensive Margin](#)
  - 5.5.2. [Sociodemographic Disparities](#)
  - 5.5.3. [Substitution of Time and Effort Towards Measured Aspects of Care](#)
- 5.6. [Scope of Testing for Unintended Consequences](#)

## 5.1. Definition

*Unintended consequences* of PCPI performance measures are the unforeseen effects of measurement on processes of care, patient outcomes, and/or the functioning of the larger healthcare system. In the context of evaluating the PCPI performance measures, unintended consequences can be interpreted to refer to any effects of measurement other than those directly achieving the purposes of performance measurement as specified by the American Medical Association (AMA), the Joint Commission (TJC), and the National Committee for Quality Assurance (NCQA) in their Consensus Statement on Principles for Performance Measurement in Health Care. According to the Consensus Statement, performance measures are to:<sup>28</sup>

- *Provide a quantitative basis for physicians, provider organizations and managed care plans to continuously improve outcomes and the care processes through which those outcomes are achieved;*
- *Provide information needed for quality oversight by regulating bodies, including regulatory agencies and private sector accrediting bodies;*
- *Provide comparative information to assist consumers and purchasers, both public and private, in selecting among provider organizations and health plans; and*
- *Facilitate prudent management of healthcare resources.*

## 5.2. Types of Unintended Consequences Arising from the Use of Performance Measures in Reporting and/or Incentive Programs

Although adverse unintended effects of performance measurement in reporting and/or pay-for-performance programs have received attention in extant literature,<sup>29</sup> there may also be positive unintended effects. The broad definition of unintended consequences given in Section 5.1. encompasses both possibilities, whether such consequences are manifest at the “micro”-level (ie, at the level of individual physicians, individual patients, physician-patient dyads), “meso”-level (ie, practice organizations, plans, networks), or “macro”-level (ie, healthcare markets or higher aggregate units such as states).

### 5.2.1. *Unintended consequences at the physician/physician-patient level (“micro”-level effects)*

Several unintended consequences of performance measurement within reporting and/or pay-for-performance programs have been noted in existing literature.

#### 5.2.1.1. *Unintentional penalization of physicians due to case mix*

If measures used in a reporting and/or incentive program are inadequately risk-adjusted, and assuming physicians act as perfect agents for their patients (the “principals”), then physicians

---

<sup>28</sup> Physician Consortium for Performance Improvement. Principles for Performance Measurement in Healthcare: A Consensus Statement from The American Medical Association and The Joint Commission on Accreditation of Healthcare Organizations and The National Committee for Quality Assurance.

<sup>29</sup> (See for example) Werner RM, Asch DA. Clinical concerns about clinical performance measurement. *Ann Fam Med.* 2007;5:159-163.

who treat disproportionate shares of patients whose clinical and/or sociodemographic profile present unfavorable risks with respect to performance measurement may be “penalized.”<sup>30,31</sup>

### 5.2.1.2. Patient selection

Each physician makes a judgment as to the selection of the range of patients that he or she accepts into his or her practice, as well as which patients the physician will choose to offer a certain treatment. If measures used in a reporting and/or incentive program are inadequately risk-adjusted, and if physicians act as *imperfect* agents for their patients, then physicians may face incentives to engage in strategic selection of the range of patients a physician accepts into their practice or the range of patients a physician offers a specific treatment.<sup>32</sup>

Physicians may face incentives to attract healthier patients and to create barriers to avoid or drop patients whose clinical and/or sociodemographic profile present unfavorable risks (ie, ‘cream’ or ‘dump’).<sup>33</sup>

Physicians may face incentives to inappropriately apply exception criteria to exclude certain patients from the denominators of performance measures (in other words, restrict the range of patients to whom they provide the measured service, or restrict the range of patients reported).<sup>34,35</sup>

### 5.2.1.3. Allocation of effort and choice of services provided

Patient care inherently involves multi-tasking. If componential tasks are differentially evaluated, then physicians will face incentives to allocate their effort accordingly even if it results in an allocation of effort that may be suboptimal given the clinical needs and preferences of patients.<sup>36,37</sup> If some activities are subject to measurement, and others are not, there will be incentives for physicians to put greater effort into performing those activities that are measured. Subject to some constraint – for example, time constraints during patient visits – physicians may choose to perform those activities that are measured at the expense of those activities that are not measured. In other words, performance measurement may distort priorities in clinical encounters, leading to a prioritization that maximizes physician utility (ie, fulfills performance measures)

---

<sup>30</sup> Epstein AM, Lee TH, Hamel MB. Paying physicians for high-quality care. *NEJM*. 2004;350:406-410.

<sup>31</sup> Zaslavsky AM, Hochheimer JN, Schneider EC et al. Impact of sociodemographic case mix on the HEDIS measures of health plan quality. *Med Care*. 2000;38:981-992.

<sup>32</sup> For an exposition on the principal-agent problem in physician-patient relationships, see: McGuire TG. Chapter 9 Physician agency. In: Anthony J. Culyer and Joseph P. Newhouse, Editor(s), *Handbook of Health Economics*, Elsevier, 2000, Volume 1, Part 1, Pages 461-536.

<sup>33</sup> This concern shares theoretical similarities with the model of patient selection here: Ellis RP. Creaming, skimping and dumping: provider competition on the intensive and extensive margins. *Journal of Health Economics*. 1998;17(5):537-555.

<sup>34</sup> C.f. Doran T, Fullwood C, Gravelle H, Reeves D, Kontopantelis E, Hiroeh U, Roland M. Pay-for-performance programs in family practices in the United Kingdom. *NEJM*. 2006;355(4):375-84.

<sup>35</sup> Gravelle H, Sutton M, Ma A. Paying for quality: British general practitioners’ performance under the quality and outcomes framework. 2006.

<sup>36</sup> This issue is identical to the problem of multi-tasking in economics. See: Holmstrom B, Milgrom P. Multitask principal-agent analyses: incentive contracts, asset ownership and job design. *Journal of Law, Economics and Organization*. 1991;7:24-51.

<sup>37</sup> An comprehensive overview of the economic literature on the problem of evaluating/rewarding agents on multi-component tasks can be found here: Prendergast C. The provision of incentives in firms. *Journal of Economic Literature*. 1999;37(1):7-63.

rather than one that maximizes patient utility.<sup>38</sup> The risk to quality is that performance evaluation may paradoxically lead physicians to override patient preferences and choice in favor of measure-centered care as opposed to patient-centered care.

Where multi-tasking is subject to differential rewards, whether reputational and/or financial, physicians may face incentives to inappropriately provide services defensively, or to inappropriately provide services to strategically inflate numerators.<sup>39</sup> Performance measurement in reporting and/or incentive programs may thus paradoxically increase rates of misuse or overuse in some circumstances.

### ***5.2.2. Unintended consequences at the organizational level (“meso”-level effects)***

#### ***5.2.2.1. Disruptions in clinical and organizational workflow***

Data collection and reporting requirements associated with performance measurement may disrupt clinical workflow patterns of physicians and support staff in ways that paradoxically detract from overall quality of care or worse, place patients at greater risk of danger.<sup>40</sup> Performance measurement similarly creates extra demand for data collection and documentation regardless of measurement modality, and the workflow changes associated with measurement may carry have potential to cause adverse events.

### ***5.2.3. Unintended consequences at the aggregate level (“macro”-level effects)***

Unintended consequences may affect physician-patient relationships and patient-physician practice relationships. These consequences may aggregate up to the level of health service markets and beyond.

Incentives for physicians to employ strategic patient selection processes in choosing the patients they will treat may, at the aggregate level, perpetuate and/or exacerbate disparities in care.<sup>41</sup>

Disparities may be perpetuated or exacerbated through a second mechanism. To the extent that information on physician performance is used by consumers (ie, patients) in healthcare decision-making, extant socioeconomic disparities – particularly education differentials and access to technology --, may create parallel differentials in the healthcare sector. Patients with more resources may be better able to gain the full advantage of performance information: they may be more likely to know about and access the physician performance information; they may be better able to understand the information; and they may be subject to fewer transportation, financial and other constraints that may otherwise limit their effective choice set of physicians.

Unintended consequences are not to be confused with artifacts of performance measures and specifications whereby flaws in technical implementation and integration of measures within information systems, and/or measure misspecifications or under-specification lead to unintended

---

<sup>38</sup> Werner RM, Asch DA. Clinical concerns about clinical performance measurement. *Ann Fam Med*. 2007;5:159-163.

<sup>39</sup> Wachter RM. Expected and unanticipated consequences of the quality and information technology revolutions. *JAMA*. 2006;295(23):2780-2783.

<sup>40</sup> Wachter RM. Expected and unanticipated consequences of the quality and information technology revolutions. *JAMA*. 2006;295(23):2780-2783.

<sup>41</sup> Casalino LP, Elster A. Will pay-for-performance and quality reporting affect health care disparities? *Health Affairs*. 2007;26(3):w405-214.

exception of certain patients or measurement error. Investigation of these artifacts are discussed in Section II (Feasibility and Implementation).

#### **5.2.4. Positive unintended consequences**

Positive unintended consequences have been much less explored in current discussions of performance measurement. Yet, it is not inconceivable that implementing performance measurement within a specific clinical domain could have positive spillover effects on the quality of care in other clinical domains, or for other patient populations other than those targeted by a specific measure or measure set. These positive “spillovers” occur when the marginal social benefit of performance measurement exceeds the marginal (private) benefit accruing to one patient or one group of patients who are the direct beneficiaries of a particular measure or measure set. For example, positive unintended consequences could occur if performance measurement in one clinical measurement area sets into motion changes in physician behavior and/or changes in the organization and structure of a physician’s practice that leads to improvements in other care.

Positive externalities may spill beyond the boundaries of a given practice. Performance measurement and quality improvement implemented in one practice may diffuse to other practices, particularly in competitive environments as other practices strive to maintain market share and position.<sup>42,43,44</sup>

### **5.3. Factors Influencing the Occurrence of Unintended Consequences**

The potential unintended consequences associated with performance measurement in reporting and/or incentive programs is not unique to performance measurement, nor should it be assumed that these consequences will materialize simply because the potential exists.<sup>45</sup> The extent to which unintended consequences are realized may depend on any one or combination of a variety of conditions:

- Where physician performance is reported outside the practice to payers, or to the general public, the degree of competition perceived by physicians (ie, perceived threat to market share) may intensify incentives to engage in strategic patient selection and/or re-prioritize care to ensure that performance measures are fulfilled.
- The magnitude of reputational or financial rewards or consequences for physician performance may intensify incentives for physicians to engage in strategic patient selection behavior. The risk of unintended consequences resulting from strategic physician behavior may be greater in “high stakes” situations: for example, when performance measurement is linked to re-certification or re-credentialing; (de-)

---

<sup>42</sup> Ferris TG. Improving quality improvement research. *Effective Clinical Practice*. 2000;3:40-44.

<sup>43</sup> Several examples of positive spillovers from quality improvement initiatives/activities are described here: the Business Roundtable, *The Spillover Effect: How Quality Improvement Efforts by Large Employers Benefit Health Care in the Community*. (June 1998).

<sup>44</sup> For a review of mechanisms through which positive spillovers may occur, c.f. Baker L. Managed care spillover effects. *Annual Review of Public Health*. 2003;24:435-456.

<sup>45</sup> c.f. Holmstrom B, Milgrom P. Multitask principal-agent analyses: incentive contracts, asset ownership and job design. *Journal of Law, Economics and Organization*. 1991;7:24-51.

selection from managed care plans or other networks; financial reward programs; and large scale public reporting programs.<sup>46</sup>

- The probability that “gaming”(the inappropriate exclusion from the measure of patients for whom the targets have been missed<sup>47</sup>) will be detected and exposed, will also influence the likelihood of those unintended consequences that result from strategic physician behavior.
- If physicians act as perfect agents for their patients, the unintended consequences resulting from strategic physician behavior are reduced.

## 5.4. Unintended Consequences and PCPI Performance Measures

### 5.4.1. Recommendation

Empirical investigations should be undertaken to describe the prevalence and patterns of unintended consequences arising from the use of PCPI performance measures within reporting and/or incentive systems. Researchers and consumers of research on unintended consequences should bear in mind the analytic separation of performance measures from their use in reporting and/or incentive programs. Researchers reporting on unintended consequences should take care to consider the above factors influencing the prevalence and patterns of unintended consequences.

**Recommendation 5-1. Empirical investigations should be undertaken to analyze the prevalence and patterns of unintended consequences arising from the implementation and use of PCPI performance measures in reporting and/or incentive systems.**

### 5.4.2. Rationale

Unintended consequences are not inherent in performance measures *per se*; rather, they arise from the coupling of performance measurement to reporting systems and/or incentive systems (such as pay-for-performance structures and/or reputational rating/ranking systems). The structure of reporting and incentive systems, as well as the manner within which PCPI measures are implemented, may vary substantially across healthcare settings and quality improvement initiatives, and it may not be within the immediate purview of the PCPI to specify acceptable designs of reporting and/or incentive systems for use with PCPI measures. In short, unintended consequences do not signal an intrinsic flaw with the design and/or specification of PCPI measures themselves; however, there may be ways that measures and their specifications can be further refined to mitigate the potential for inappropriate manipulation of performance measures. For this reason, the PCPI MIE recommends that studies be carried out to investigate the prevalence and nature of unintended consequences associated with the use of PCPI performance measures.

<sup>46</sup> c.f. Jacob B. Accountability, incentives and behavior: evidence from school reform in Chicago. *Journal of Public Economics*. 2005;89:761-796.

<sup>47</sup> Doran T, Fullwood C, Gravelle H, et al. Pay-for-performance programs in family practices in the United Kingdom. *N Engl J Med*. 2006;355:375–384.+

## **5.5. General Considerations for Investigating Unintended Consequences**

### ***5.5.1. Strategic Patient Selection***

Studies of strategic selection of patients at the extensive margin may examine differences in the prevalence of patients meeting the exception criteria for a given PCPI measure before and after the implementation of a PCPI measure. Significant differences in the prevalence of patients meeting exception criteria may suggest potential gaming, but additional investigation will be necessary to rule out legitimate reasons for the difference(s).

An alternative approach to investigate patient selection is to audit reported exceptions. A high rate of invalid exceptions may suggest strategic selection of patients. However, auditing exceptions for purposes of studying unintended consequences should be carried out only after there is reasonable confidence that discrepancies are not attributable to technical difficulties or problems in the implementation and integration of a performance measure within a practice's health information system (see Section II).

### ***5.5.2. Sociodemographic Disparities***

To explore sociodemographic disparities within practices, subgroup analyses of exceptions (see above) and outcomes (see Section III) are encouraged. However, it is unclear if strategic patient selection occurs between patients (within practice) or whether it is primarily manifest between practices where practices serving a disproportionate share of sociodemographically "risky" patients engage in selection. This question is best addressed in studies involving large numbers of practices and patients; however, the costliness of these studies prohibits the MIE from prescribing them as a standard component of testing PCPI measures.

### ***5.5.3. Substitution of Time and Effort towards Measured Aspects of Care***

It is difficult to specify research recommendations or standards for investigating the concern that physicians will re-prioritize services and aspects of care during visits to maximize performance measures rather than patient utility. However, an indirect approach to addressing this question that can be broadly applicable across settings and measure sets is to compare patient satisfaction as measured through standardized survey instruments (such as CAHPS) before performance measurement and after performance measurement. Significant differences may be suggestive of possible re-prioritization or other changes in the visit content and style due to performance measurement that lower patient utility.

## **5.6. Scope of Testing for Unintended Consequences**

The prevalence and patterns of unintended consequences associated with the implementation and/or use of PCPI performance measures in reporting and/or incentive systems should be evaluated in both private and public ambulatory and/or inpatient healthcare delivery settings as appropriate to the measure (set) under consideration.

## SECTION VI. APPLICATIONS

### Section VI Outline

- 6.1. [Applications of PCPI Performance Measures](#)
- 6.2. [Studies Investigating the Association between Performance and Outcomes](#)
  - 6.2.1. Recommendation
  - 6.2.2. Rationale
- 6.3. [Studies Investigating the Effectiveness of Performance Measures in Quality Improvement](#)
  - 6.3.1. Recommendation
  - 6.3.2. Rationale

## 6.1. Applications of PCPI Performance Measures

The PCPI MIE feels that it is necessary to acknowledge the demand for applied studies of PCPI performance measures in quality improvement. Such demand has expressed in two ways:

- A demand for studies investigating the correlation between performance on PCPI measures and patient outcomes
- A demand for studies investigating the effectiveness of using PCPI performance measures in quality improvement interventions to improve processes of care

## 6.2. Studies Investigating the Association between Performance and Outcomes

### 6.2.1. Recommendation

The association between performance on PCPI measures and patient outcomes concerns the predictive validity of a PCPI measure. Thus, the PCPI MIE refers to Recommendation 4-4 and recommends that such studies be considered optional but desirable for PCPI measures that focus on processes of care that are established on a strong, conclusive evidence base. Studies of predictive validity are to be compulsory for all PCPI measures focusing on processes of care that are not established on a clear, strong and conclusive evidence base<sup>48</sup>.

**Recommendation 6-1 (Recommendation 4-4). Predictive validity testing is recommended where possible for all PCPI measures focusing on processes of care that are not supported by a strong evidence base. Predictive validity testing is of value but dependant upon data availability and the stage of implementation for all PCPI measures focusing on processes of care that are founded upon a strong, conclusive evidence base.**

### 6.2.2. Rationale

Please refer to Section 4.5.3.

## 6.3. Studies Investigating the Effectiveness of Performance Measures in Quality Improvement

### 6.3.1. Recommendation

The PCPI MIE recommends that applied studies of PCPI performance measures in quality improvement interventions be considered *nonessential*, but *supplemental* material in the evidence base for a PCPI performance measure.

---

<sup>48</sup> Physician Consortium for Performance Improvement® (PCPI) Position Statement. The Evidence Base Required for Measures Development. Approved 6/26/2009. Available at: <http://www.ama-assn.org/ama1/pub/upload/mm/370/pcpi-evidence-based-statement.pdf>

- The PCPI MIE recommends that PCPI Measure Development Work Groups review studies of the use of PCPI measures in quality improvement interventions or programs as such research may provide implications that inform the refinement of measure definitions and/or specifications.

***Recommendation 6-2. Findings from applied studies of the use of PCPI measures in quality improvement interventions/programs should not be considered admissible evidence regarding the scientific soundness of a PCPI measure or measure set. However, applied studies should be reviewed by PCPI Measure Development Work Groups for possible implications that may inform further refinement of measure definitions and/or specifications.***

### ***6.3.2. Rationale***

The PCPI MIE cautions against accepting findings from applied research such as trials of quality improvement interventions or programs as evidence in support of, or against the scientific soundness and/or feasibility of PCPI performance measures. The PCPI measures are developed as instruments for measurement, and the scientific evaluation of PCPI measures by the PCPI and other interested stakeholders should focus on functional capabilities of the PCPI for measurement. The PCPI MIE regards performance measurement as a necessary component of quality improvement research, but takes the position that changes in quality cannot be attributable to characteristics intrinsic to the measure itself, but depends on the manner in which measures are used, as well as the design, effectiveness, and execution of the intervention or program within which PCPI measures are applied. The PCPI does not make any prescriptions regarding interventional approach or modality. Differential effects of quality improvement interventions that use PCPI measures may depend on the differences in subjects across sites, the nature of the intervention (eg, physician-centered traditional continued medical education (CME) versus a practice systems-approach), the intensity or duration of the intervention, and/or a number of other study factors that are not within the purview of the Consortium or its Measure Development Work Groups.

The PCPI MIE does recognize, however, that applied studies may serve as examples for the use of PCPI measures in quality improvement activities, and that such studies may provide implications (direct or indirect) that may inform future refinement of measure definition or specification. Therefore, the PCPI MIE wishes to encourage the review of applied studies by PCPI Measure Development Work Groups.

## **APPENDICES**

## **APPENDIX A: LIST OF KEY RECOMMENDATIONS**

**Recommendation.** All PCPI performance measures should be tested in each of the following testing areas:

- Needs Assessment (Testing Area 1)
- Feasibility (Testing Area 2)
- Reliability (Testing Area 3)
- Validity (Testing Area 4)
- Unintended Consequences (Testing Area 5)
- Applications (Testing Area 6)

**Recommendation 1-1.** For existing performance measures, it is recommended that the existence of a gap in care be re-evaluated by the measure developing body or other evaluators at the time that the measurement set is being formally updated (for the PCPI, measures are updated every 3 years).

**Recommendation 1-2.** For an existing PCPI measure, acceptable types of evidence for documenting significant gaps in care include: 1) previously published studies in peer-reviewed publications; and 2) secondary analysis of existing data.

**Recommendation 2-1.** At least one feasibility and implementation study should accompany all PCPI performance measures. Studies should include: (a) a description of the implementation strategy; (b) feasibility analysis of data collection; (c) barriers analysis; and (d) an analysis of resource utilization/costs.

**Recommendation 2-2.** Feasibility of implementing PCPI performance measures should be demonstrated in the following settings: (a) clinical practices using EHR-based measurement modalities; (b) clinical practices using administrative/claims data-based measurement modalities; and (c) clinical practices using paper medical record data-based measurement modalities.

**Recommendation 2-3.** Feasibility of implementing PCPI performance measures should be demonstrated in private and public clinical environments.

**Recommendation 3-1.** All PCPI performance measures should undergo reliability testing in all measurement modalities for which technical specifications are developed: EHR- or registry-based measurement, administrative/claims-based measurement, registry-based measurement and paper medical record data-based measurement.

**Recommendation 3-2.** All PCPI performance measures should undergo inter-abstractor reliability testing to evaluate medical record data-based measurement modalities.

**Recommendation 3-3.** All PCPI performance measures should undergo the following tests of parallel forms reliability: (a) EHR-based measurement vs. manual review; (b) administrative/claims-based measurement vs. manual review; (c) registry-based measurement vs. manual review.

**Recommendation 3-4.** Assessments of test-retest reliability for PCPI performance measures are not recommended.

**Recommendation 3-5.** Assessments of internal consistency for PCPI performance measures are not recommended.

**Recommendation 3-6.** When feasible, it is strongly recommended that the reliability of each PCPI measure should be formally evaluated in: (a) private practice settings; (b) publicly funded (federal/state/county/municipal) settings that serve disproportionate shares of clinically and/or sociodemographically vulnerable patients; (c) practices that use EHR-based performance measurement modalities; (d) practices that use administrative/claims data-based measurement modalities; and (e) practices that use paper medical record data-based measurement modalities.

**Recommendation 4-1.** All PCPI performance measures are assessed for face validity by expert Work Group members during the development process. No additional testing to establish face validity is required, unless requested by the workgroup.

**Recommendation 4-2.** All PCPI performance measures are assessed for content validity by expert Work Group members during the development process. No additional testing to establish content validity is required, unless requested by the workgroup.

**Recommendation 4-3.** All individual PCPI performance measures are exempt from construct validity testing.

**Recommendation 4-4.** Predictive validity testing is recommended where possible for all PCPI measures focusing on processes of care that are not supported by a strong evidence base. Predictive validity testing is of value but dependant upon data availability and the stage of implementation for all PCPI measures focusing on processes of care that are founded upon a strong, conclusive evidence base.

**Recommendation 4-5.** It is desirable for PCPI measures to demonstrate predictive validity and/or measurement level validity studies in diverse settings, including: (a) private practice settings; (b) publicly funded (federal/state/county/municipal) settings that serve disproportionate shares of clinically and/or sociodemographically vulnerable patients; (c) practices that use EHR-based performance measurement modalities; (d) practices that use administrative/claims data-based measurement modalities; and (e) practices that use paper medical record data-based measurement modalities.

**Recommendation 5-1.** Empirical investigations should be undertaken to analyze the prevalence and patterns of unintended consequences arising from the implementation and use of PCPI performance measures in reporting and/or incentive systems.

**Recommendation 6-1.** (Recommendation 4-4). Predictive validity testing is recommended where possible for all PCPI measures focusing on processes of care that are not supported by a strong evidence base. Predictive validity testing is of value but dependant upon data availability and the stage of implementation for all PCPI measures focusing on processes of care that are founded upon a strong, conclusive evidence base.

**Recommendation 6-2.** Findings from applied studies of the use of PCPI measures in quality improvement interventions/programs should not be considered admissible evidence regarding the scientific soundness of a PCPI measure or measure set. However, applied studies should be reviewed by PCPI Measure Development Work Groups for possible implications that may inform further refinement of measure definitions and/or specifications.

**APPENDIX B: LIST OF EXISTING PCPI MEASURES (as of Sept 1, 2010)**

**42 measure sets, 270 Performance Measures**  
**[www.physicianconsortium.org](http://www.physicianconsortium.org)**

Descriptions and specifications for PCPI performance measures are available for 42 clinical topics and conditions. A list of 270 PCPI performance measures is available for implementation.

Link to Measures and Specifications:

<http://www.ama-assn.org/ama/pub/physician-resources/clinical-practice-improvement/clinical-quality/physician-consortium-performance-improvement/pcpi-measures.shtml>

## **APPENDIX C: EXAMPLE PUBLISHED TESTING EFFORTS FOR PCPI MEASURES**

- **Backus LI, Boothroyd DB, Phillips BR, Belperio PS. National Quality Forum Performance Measures. Arch Intern Med. 2010;170(14):1239-1246**  
National performance rates on HIV measures were generally high, but variation in rates across facilities revealed room for improvement. Both patient and resource factors had an impact on the likelihood of receipt of indicated care.
- **Baker DW, Persell SD, Thompson JA, Soman NS, Burgner KM, Liss BA, Kmetik KS. Automated Review of Electronic Health Records to Assess Quality of Care for Outpatients with Heart Failure. Ann Intern Med. 2007;146:270-277.**  
Automated review of EHR data to measure the quality of care of outpatients with heart failure missed many exclusion criteria for medications documented only in providers' notes. As a result, it sometimes underestimated performance on medication-based quality measures.
- **Linder JA, Kaleba EO, Kmetik KS. Using Electronic Health Records to Measure Physician Performance for Acute Conditions in Primary Care Empirical Evaluation of the CAP Clinical Quality Measure Set. Medical Care. 2009;47:208–216.**  
Although EHRs offer potential advantages for performance measurement for acute conditions, accurate identification of pneumonia visits was challenging, performance generally appeared poor, and much of the data were not in coded form.
- **Patel MM, Eisenberg L, Witsell D, Schulz KA. Assessment of acute otitis externa and otitis media with effusion performance measures in otolaryngology practices. Otolaryngology-Head and Neck Surgery. 2008;139:490-494.**  
Although compliance in this study was generally high across both measure sets, actual use of the face sheet forms for appropriate patients was lower than the 80% reporting mandate by the Centers for Medicare and Medicaid Services that allows physicians to receive the monetary bonus. Incentive-based reporting should be continuously investigated to assess challenges for evaluating current measures.
- **Persell SD, Kho AN, Thompson JA, Baker DW. Improving Hypertension Quality Measurement Using Electronic Health Records. Medical Care. 2009;47(4):388-394.**  
It is possible to use electronic health record data to devise hypertension measures that may better reflect who has actionable uncontrolled blood pressure, do not penalize clinicians treating resistant hypertension patients, reduce the encouragement of potentially unsafe practices, and identify patients possibly receiving poor care with no hypertension diagnosis. This could improve the detection of true quality problems and remove incentives to over treat or stop caring for patients with resistant hypertension.