

## UnitedHealthcare Comments on NQF Resource Use White Paper

1. We applaud the NQF for developing a white paper on the important topic of resource use measurement, and we appreciate the opportunity to comment.
2. We agree with the formulation that resource use measurement must be connected to quality measurement. One example of connecting quality to cost assessment is our UnitedHealth Premium® designation program, in which only physicians designated for quality of care are eligible for designation as cost-efficient.
3. The white paper indicates that NQF would favor if not require resource measures that are connected directly to one or more NQF Endorsed® quality measures.
4. We are concerned that this would create too narrow a set of resource use measures
  - a. We note that the current NQF Endorsed measures have gaps in terms of appropriateness measures, outcome measures, and in general in the area of specialty and surgical care.
  - b. Resource use assessment in those areas is necessary as a part of addressing the urgent affordability agenda.
  - c. Conversely, the framework proposed in the white paper underscores the need for measures that fill gaps in the NQF Endorsed measure set
5. A requirement for corresponding NQF Endorsed quality measures could result in resource use measures limited to narrow scopes (i.e. specific types or settings of resource use).
6. On the other hand, episode of care techniques are commonly to create composite resource measures, as the white paper appears to anticipate in pages 39-41, for example.
7. It is unclear therefore how the NQF proposed formulation would allow creation of NQF Endorsed composite episode resource measures
8. In work at UHC and elsewhere, we have found that creating composite episode measures and then decomposing them into key cost drivers enables the identification of overuse, for example repeated complex imaging, polypharmacy, antibiotic overuse, and inappropriate site of service use.
  - a. Overuse is per se a quality issue
  - b. Therefore in these situations the requirement of matching NQF Endorsed quality measures would inhibit use of composite methods for addressing such overuse issues
9. We recommend that NQF split the topic composite episode measures out of this paper and into a separate white paper to be developed at a later time.
  - a. The UnitedHealth Premium® physician designation program (UHPD), soon to be in its fifth release, has used composite episode measures from the start.
  - b. Given our more than 6 years' experience, we can state unequivocally that these measures are inherently complex, requiring multiple levels of methodological decisions.
  - c. The methodology for UHPD is publicly available at [www.unitedhealthcareonline.com](http://www.unitedhealthcareonline.com):

[https://www.unitedhealthcareonline.com/ccmcontent/ProviderII/UHC/en-US/Assets/ProviderStaticFiles/ProviderStaticFilesPdf/Clinician%20Resources/UnitedHealth%20Premium%20Designation%20Program/UnitedHealth\\_Premium\\_Methodology\\_Detailed\\_Summer\\_2010.pdf](https://www.unitedhealthcareonline.com/ccmcontent/ProviderII/UHC/en-US/Assets/ProviderStaticFiles/ProviderStaticFilesPdf/Clinician%20Resources/UnitedHealth%20Premium%20Designation%20Program/UnitedHealth_Premium_Methodology_Detailed_Summer_2010.pdf)

- d. The sections describing UHPD cost efficiency methodology total approximately 20 pages, single spaced (i.e. the same length as the entire NQF white paper).
  - e. The start of the art of composite episode measures is evolving rapidly.
  - f. Recent papers have reinforced the need for statistical testing. Parametric tests have been used for many years but non-parametric tests may be preferred and have been developed at UnitedHealthcare.
  - g. For all these reasons we respectfully submit that the discussion of composite episode measures should
    - i. Be separated into a separate white paper
    - ii. Receive much more extensive treatment
    - iii. Contain input from those with practical experience creating composite episode measures, including the UHPD system
    - iv. Reflect and respect the evolving nature of this area of performance measurement
10. In regards to NQF's multiple types of resource use measures, UHC agrees that measures or combinations of measures should be selected in light of the goal of measurement. In particular, we agree that population measures should supplement episode measures, and we are working towards that end.
11. In regards to limitations of claims and administrative data, the white paper does not present a balanced view.
- a. Sample size is not a limitation specific to claims or administrative data, in fact sample sizes are often larger when drawn from administrative data than chart data due to the high cost of chart abstraction.
  - b. Similarly, "black box" methodology can be applied to many types of data
  - c. Therefore both of these comments should be removed from the section on limitations of claims data, and they should reflect that all data is subject to those potential issues.
  - d. In addition to EMR data, ICD-10 will increase the utility of claims and administrative data.
  - e. To be balanced, the report should also mention that chart data itself has been shown to contain errors when compared to a video tape of the actual encounter, and that chart data is inherently incomplete because it reflects only that data developed or collect by one physician.
12. Generally we agree with the principles for measures of resource use, however they could be improved in several ways:
- a. The principles need to be stated in clearly defined terms. For example, what are comprehensive measures in this case, and what are they preferable too? Depending on the goals of measurement they may not always be preferable
  - b. In the cases where comprehensive or composite measures are indeed preferable, they will be inherently complex. That implies a need to strike a

balance with the principle that results should be simple and readily interpretable. The language of the latter principle should be adjusted to reflect the need for that balance

- c. Rigid interpretations of principles will make them harder to meet. For example, benchmark construction can be made transparent, but in many cases the individual data underlying the resulting benchmark value cannot (for example, due to contractual limitations on disclosure of fee schedules). The principles should include language about reasonability to reflect the potential for such issues.
13. Three principles should be added to those outlined
- a. Connection to the Triple Aim and the NPP: Resource use measures, and in addition appropriate use, are explicitly implied in both the Triple Aim of Better Care, Better Health and Lower Costs, as well as in the National Priority Partnership (which is co-sponsored by NQF) priority of eliminating overuse
  - b. Importance of measures of appropriateness and overuse: The principles should reflect that measures intended to increase appropriate use or decrease overuse inherently increase efficiency (they decrease cost and increase quality through avoidance of risk or side effects).
  - c. Velocity of implementation: Quality measures have been in place for more than 20 years but resource use measures are still few and far between. Due to soaring health care costs and the affordability crisis, there is an urgent need for increased efficiency. Therefore there must be principles addressing velocity of development such as the following:
    - i. Simple measures that affect important areas of inefficiency are important [and at least as important as comprehensive measures]
    - ii. If two measures address the same area, the measure that can be brought to bear more quickly is preferred
14. Under Evaluation Criteria, we do not believe “Scientific Acceptability” best describes the criterion for “the extent to which the measures, as specified, produce consistent (reliable) and accurate (valid) results about the cost or resources used to deliver care.” Scientific acceptability may apply to quality measures based on science (evidence-based medicine). We agree with the attributes of resource use as described, however they do not reflect scientific acceptability. We suggest describing these simply as “Acceptable Measure Properties.”
15. Two technical issues
- a. Reliability: the white paper follows standard English usage of the word, reliability. Given recent publicity, the audience of the white paper will surely be familiar with the RAND research on *statistical* reliability. We recommend that the NQF white paper carefully and clearly defines its meaning of reliability the first time the term is used, and again when necessary, as well as explicitly distinguish it from statistical reliability.
  - b. Construction of O:E ratio: The discussion on pp 39 and 40 regarding different methods of O:E construction is misleading. The two methods (mean of O:E vs total O: total E) do indeed yield different results, but the mean O:E is not methodologically sound. Composite O:E ratios must preserve case mix and severity adjustment (and in fact are constructed so as to do that through

indirect standardization). The mean O:E method does not do that. See embedded Excel file for counterexamples. Therefore the fact that the mean O:E method yields a different result is not relevant.

16. We commend NQF for noting that the complexity of statistical tests should be balanced with the need for transparency and interpretability (line 1259). We suggest these warnings be moved up to the front of the paper, and be repeated where necessary when discussing advanced statistical techniques (for example, lines 1166-1167 discussing multi-level regression and Monte Carlo simulation).
17. NQF mentions in passing issues related to statistical reliability (line 1348 and reference). UHC disagrees with the use of statistical reliability in performance measurement. It does not reflect the magnitude of performance difference from benchmark, and therefore is insufficient for performance measurement. Statistical reliability is not a familiar concept outside statistical circles, and the methods themselves are abstruse. Thus it fails the transparency and interpretability test

We prefer tests of statistical confidence. Because the field is evolving, and also because there are many choices of statistical tests, we ask NQF to couch its recommendations in general terms rather than prescribing specific tests or methods.

**Demonstration that "Mean O:E" is not methodologically sound**

**Mean O:E methodology**

**Physician 1**

	Physician's Episode Count	Physician's Total Cost (O)	Cost per Episode	Peer Group Cost per Episode	Total Expected Cost (E)	O:E for Condition
Calculation:	B	C	= C / B	E	= B x E	= C / F
Condition A	10	\$ 1,200	\$ 120	\$ 100	\$ 1,000	<b>1.2</b>
Condition B	2	\$ 800	\$ 400	\$ 500	\$ 1,000	<b>0.8</b>

**Physician 2**

(In which all parameters are the same except case mix)

	Physician's Episode Count	Physician's Total Cost (O)	Cost per Episode	Peer Group Cost per Episode	Total Expected Cost (E)	O:E for Condition
Calculation:	B	C	= C / B	E	= B x E	= C / F
Condition A	2	\$ 240	\$ 120	\$ 100	\$ 200	<b>1.2</b>
Condition B	10	\$ 4,000	\$ 400	\$ 500	\$ 5,000	<b>0.8</b>

If 1.2 and 0.8 are averaged, both physicians have the same result, "Mean O:E" of 1.0

Note however that this methodology washes out case mix adjustment. If the two rows had been same condition but different severity, it would wash out severity adjustment. Therefore this methodology is not sound, because the result should reflect case-mix and severity adjustments

**Usual O:E methodology:**

**Physician 1**

	Physician's Episode Count	Physician's Total Cost (O)	Cost per Episode	Peer Group Cost per Episode	Total Expected Cost (E)	Total O to Total E
Calculation:	B	C	= C / B	E	= B x E	= ΣC / ΣF
Condition A	10	\$ 1,200	\$ 120	\$ 100	\$ 1,000	
Condition B	2	\$ 800	\$ 400	\$ 500	\$ 1,000	
<b>Total:</b>		<b>\$ 2,000</b>			<b>\$ 2,000</b>	<b>1.00</b>

**Physician 2**

(In which all parameters are the same except case mix)

	Physician's Episode Count	Physician's Total Cost (O)	Cost per Episode	Peer Group Cost per Episode	Total Expected Cost (E)	Total O to Total E
Calculation:	B	C	= C / B	E	= B x E	= ΣC / ΣF
Condition A	2	\$ 240	\$ 120	\$ 100	\$ 200	
Condition B	10	\$ 4,000	\$ 400	\$ 500	\$ 5,000	
<b>Total:</b>		<b>\$ 4,240</b>			<b>\$ 5,200</b>	<b>0.82</b>

The usual O:E methodology accounts for differences in case mix. If instead of two conditions, the rows were the same condition but different severity (e.g. "Bronchitis without complication" and "Bronchitis with complication") the severity differences would be preserved.