

Breaking Narrative Ground: Innovative Methods for Rigorously Eliciting and Assessing Patient Narratives

Rachel Grob, Mark Schlesinger, Andrew M. Parker, Dale Shaller, Lacey Rose Barre, Steven C. Martino, Melissa L. Finucane, Lise Rybowski, and Jennifer L. Cerully

Objective. To design a methodology for rigorously eliciting narratives about patients' experiences with clinical care that is potentially useful for public reporting and quality improvement.

Data Sources/Study Setting. Two rounds of experimental data ($N = 48$ each) collected in 2013–2014, using a nationally representative Internet panel.

Study Design. Our study (1) articulates and operationalizes criteria for assessing narrative elicitation protocols; (2) establishes a “gold standard” for assessment of such protocols; and (3) creates and tests a protocol for narratives about outpatient treatment experiences.

Data Collection/Extraction Methods. We randomized participants between telephone and web-based modalities and between protocols placed before and after a closed-ended survey.

Principal Findings. Elicited narratives can be assessed relative to a gold standard using four criteria: (1) meaningfulness, (2) completeness, (3) whether the narrative accurately reflects the balance of positive and negative events, and (4) representativeness, which reflects the protocol's performance across respondent subgroups. We demonstrate that a five-question protocol that has been tested and refined yields three- to sixfold increases in completeness and four- to tenfold increases in meaningfulness, compared to a single open-ended question. It performs equally well for healthy and sick patients.

Conclusions. Narrative elicitation protocols suitable for inclusion in extant patient experience surveys can be designed and tested against objective performance criteria, thus advancing the science of public reporting.

Key Words. Patient narratives, patient experiences, public reporting, consumers, qualitative methods, patient-centered care

Web-based comments written by consumers about their service providers have become ubiquitous in the United States (Vasquez 2014). Americans now consult user reviews not only when making decisions about what book to buy

or plumber to hire but also when choosing health care providers (Gao et al. 2012; Lansky 2012). Five years ago, consumer commentary about clinicians was rare, but by 2013, 31 percent of Americans had read patient reviews of health care providers online and 21 percent used them when selecting a clinician (Health Research Institute 2013). This rapid expansion contrasts with consumers' use of other clinician quality metrics, which has remained unchanged over the past decade (Kaiser Family Foundation 2015).

The proliferation of patient commentary, which is already transforming consumer behavior, could also transform public reporting of health care quality (Schlesinger et al. 2015a). When patients describe clinical encounters in their own words, their comments can substantially enhance consumer engagement with quality reports (Lagu and Lindenauer 2010; Kanouse et al. 2016). Comments also convey crucial aspects of care omitted from conventional patient experience surveys (e.g., clinicians' decision-making styles and emotional connectedness with patients) (López et al. 2010; Detz, López, and Sarkar 2013) and help consumers translate comparative quality information into coherent choices (Schlesinger et al. 2015a). It is crucial to proceed with care in generating and presenting narrative data, however, because if available comments fail to convey a full and accurate picture of interactions with clinicians, these potential benefits can be offset by choices that do not align with consumers' own needs and preferences (Trigg 2011; Coggnetta-Rieke and Guney 2014; Kanouse et al. 2016).

Patient experience surveys, such as the Consumer Assessment of Healthcare Providers and Systems (CAHPS), have demonstrated scientific rigor (Cleary et al. 2012). Accordingly, they have been endorsed by public bodies such as the National Quality Forum and incorporated into public reporting. By contrast, patient commentary on health care providers typifies dictionary definitions of "anecdotal" because they are "based on personal

Address correspondence to Mark Schlesinger, Ph.D., Professor of Health Policy, Yale University School of Public Health, Room 304 LEPH 60 College St, New Haven, CT 06520; e-mail: mark.schlesinger@yale.edu. Rachel Grob, Ph.D., M.A., is with the Center for Patient Partnerships, UW Law School, University of Wisconsin—Madison, Madison, WI; Department of Family Medicine, UW Medical School, University of Wisconsin—Madison, Madison, WI. Andrew M. Parker, Ph.D., M.S., is with the RAND Center for Decision Making under Uncertainty, Pittsburgh, PA; RAND, Pittsburgh, PA. Dale Shaller, M.P.A., is with the Shaller Consulting Group, Stillwater, MN. Lacey Rose Barre, M.P.H., is with the Department of Health Services Research, Policy, and Practice, Brown University School of Public Health, Providence, RI. Steven C. Martino, Ph.D., is with the RAND Health, Pittsburgh, PA. Melissa L. Finucane, Ph.D., and Jennifer L. Cerully, Ph.D., M.S., are with the RAND, Pittsburgh, PA. Lise Rybowski, M.B.A., is with The Severyn Group, Ashburn, VA.

experience or reported observations unverified by controlled experiments.”¹ Patient comments currently available online are anecdotal both because they are volunteered (Sick and Abraham 2011; Tompson and Wilcoxon 2014) rather than elicited from a representative sample of patients, and because they have not been collected using a methodology with proven reliability.

These limitations on currently available comments present serious challenges for the potential incorporation of narrative patient accounts into public reporting. Comments about care are volunteered by consumers in much the same way as is true for many consumer products (Vasquez 2014)—but clinical care is different in important ways from other consumer goods and services. Unlike standardized products (e.g., books or appliances) used by millions, there are typically only a few thousand patients in each clinician’s panel. If a small proportion of consumers voice their experiences, volunteered comments may be too sparse to reliably characterize each provider or practice (Gao et al. 2012). Low yield is a predictable problem with comments about experience with clinical care because Americans are less likely to comment online about health care than about other consumer experiences (Health Research Institute 2013). It is also a problem exacerbated by the reality that organizational and technological change in clinical settings that may affect patient experiences can rapidly render older comments out-of-date (Gann 2013; Greaves, Millett, and Nuki 2014).

An equally daunting difficulty with volunteered accounts is that many Americans are unsure what is reasonable to expect from their clinical care (Tu and May 2007). This is especially true for those less experienced with health care or those who have serious or complex health problems—groups of patients who are reticent to report even serious problems with care (Schlesinger 2011). If comments are missing or incomplete for these subgroups, the available narrative data will be partial and unrepresentative.

To assess the potential for moving from anecdotal to rigorously elicited patient narratives, we developed and tested a sequence of open-ended questions (hereafter, “elicitation protocol”) designed to induce reliably high-quality accounts of patients’ interactions with clinicians from a representative cross-section of the public. Our mixed methods approach was designed to treat the elicitation and reporting of narrative data with the same scientific rigor accorded closed-ended surveys (Cognetta-Rieke and Guney 2014).

This paper documents methods we employed for creating and testing elicitation protocols. The work proceeded in two stages. In the first stage, we clarified how “rigor” applies to elicited narratives, by (a) establishing a “gold

standard” of narrative expression against which “elicited patient narratives” (hereafter referred to as “EPNs”) could be compared, and (b) articulating a set of criteria for assessing protocol performance. In the second stage, we applied these criteria to create, test, and iteratively refine protocols for eliciting patient narratives about clinical care. Below, we describe methods used in each stage and then present selected findings that document the refinement of both the criteria and the final protocol. Findings presented here are intended to illustrate the methodological innovation this paper documents; full results from the elicited protocols will be presented elsewhere.

The elicitation strategy we explore here links open-ended questions to an established patient experience instrument: the CAHPS Clinician & Group Survey (CG-CAHPS). This strategy has certain benefits, including a representative sampling frame of patients who recently visited a clinician. However, our study contributes broadly to the science of public reporting by developing and explicating a methodology for rigorous elicitation of narrative patient experiences that can be adapted for *any* clinical setting or paired with *any* existing patient experience data collection instrument.

METHODS I: CRITERIA FOR ASSESSING ELICITATION PERFORMANCE

Capturing the Richness of Narrative Content

Narratives convey more nuance than quantitative ratings. As we developed a set of desired characteristics for EPNs, we proceeded with the assumption that they function not only cognitively but also at symbolic and affective levels (Greenhalgh and Hurwitz 1999; Charon 2001)—and that they require both “a teller and a listener” as they are by definition a form of discourse which requires “someone telling someone else that something happened” (Charon 2001, including quote from Smith). Narratives thus communicate information not only about what is told but also about the teller.

We drew on multiple literatures to develop criteria for assessing the effectiveness of elicitation protocols, including three bodies of research: studies of narrative medicine that focus on the diagnostic potential of narratives (Greenhalgh and Hurwitz 1999; Charon 2001); research on life stories and the nature of narrative coherence (McAdams 2006; Reese et al. 2011); and analyses of how narratives influence treatment choices (Ziebland and Wyke 2012; Shaffer, Tomek, and Hulsey 2014).

A “Gold Standard” to Measure Validity of EPNs

The validity of narratives can be assessed from a variety of perspectives, including their congruence with clinicians’ descriptions of the same encounters or with data in medical records.

Because EPNs were designed to describe experiences with clinical care *as seen through the patient’s own eyes*, the key methodological consideration we established for assessing quality of elicited data is how it measures up to a full account from that same patient. Our methodology therefore included hour-long semistructured intensive interviews with the same participants who completed the EPN. These “gold standard” interviews were expressly designed to capture nuances of experience via a set of probes, and to generate a comprehensive narrative where the “teller” was paired with an actual “listener” (the interviewer) rather than an imagined one. Identical coding techniques (described below) were applied to both EPNs and interviews. The congruence between each individual’s “gold standard” narrative from the interview and the narratives derived through EPN elicitations served as our measure of the elicitation protocol’s validity.

Defining Effective Elicitation: Criteria and Measures

The Content of EPNs: Completeness and Balance. Studies of volunteered patient anecdotes, though biased as a measure of the prevalence of concerns, are useful for identifying the categories of experiences that Americans convey in their comments (López et al. 2010; Tsianakas et al. 2012; Lagu et al. 2013), and what they say they are looking for in clinicians (Galizzi et al. 2012; Tompson and Wilcoxon 2014). Drawing on this literature and our own past surveys (Mitchell and Schlesinger 2005), we defined our criteria for analyzing EPN content as follows: (a) *completeness*, as assessed by the number of distinctive aspects of care that were described in both the EPN and the hour-long interview, and (b) *balance*, as assessed by the congruence of the relative proportion of positive versus negative experiences in the EPN and its matched interview.

Operationalizing these criteria required developing specific measures for each. For completeness, we identified 10 distinct commonly described aspects of patient–clinician interactions (see Appendix S1) and counted the number of aspects in each EPN and its corresponding interview (see coding section below). For balance, we assigned each of these identified aspects of experience to one of three valence categories: positive, negative, or neutral. The aggregate measure of balance was constructed as the ratio of positive to negative experiences conveyed in the EPN, over the ratio for the corresponding interview.

The Meaning of EPNs for the Listener: Coherence and Relevance. To specify dimensions of EPNs that render them meaningful, we drew on the literature examining how narratives convey meaning. One branch of this literature focuses on the properties of narratives that render them “coherent” for readers (McAdams 2006; Reese et al. 2011). Another examines how decision makers integrate narrative accounts with other sources of information (Winterbottom et al. 2008; Dieckmann, Slovic, and Peters 2011; Shaffer, Tomek, and Hulseay 2014). Combined, this research suggests that narratives are meaningful to others when they are both *coherent* and *relevant*—that is, told by a person somewhat similar to the listener/reader, in ways that make sense to that listener/reader (Ziebland and Wyke 2012; Shaffer and Zikmund-Fisher 2013).

We hypothesized, based on the above-cited literatures, eight attributes plausibly related to *coherence*: (1) clarity of goals (e.g., what expectations do patients have for care/clinicians?); (2) wholeness of storyline (e.g., does it have a beginning, middle, and end?); (3) chronology (e.g., is there a clear sequence of events?); (4) concreteness (e.g., are particular episodes described?); (5) texture (e.g., are experiences described in detail?); (6) emotional response (e.g., how did events feel to the patient?); (7) expressed rationale (e.g., were perceived causes of events and actions described?); and (8) consistency of evaluation (e.g., was the balance of negative and positive valence of description relatively stable throughout the narrative?). These are attributes that have been demonstrated to make real-life stories meaningful to readers or listeners rather than a set of literary standards (Schlesinger et al. 2015a).

To assess how well these eight narrative attributes derived from the literature actually fit together as a composite measure for coherence of narratives about experiences in clinical settings, we conducted an exploratory factor analysis of fully coded data (see below). Results revealed two distinct factors (see Table 1): “richness” of the storyline and specificity of the narrative. Chronology correlates with both factors. However, consistency does not closely correlate with either and is strongly negatively correlated with the second factor. We thus operationalized our primary measure of coherence for EPNs as a composite for narrative richness incorporating five component measures (rows 1–4 and 6 from Table 1; alpha coefficient of 0.83 for this composite measure of coherence). A second measure of narrative specificity is derived from concreteness and chronology (rows 6 and 7 from Table 1). As consistency did not correlate with either of these constructs, we concluded it was not applicable to patients’ narratives about clinical care.

Table 1: Dimensionality of Measures of Narrative Coherence

<i>Rotated Factor Pattern</i>	<i>Factor 1</i>	<i>Factor 2</i>
1. Texture/detail of account	0.83	-0.18
2. Clarity of rationale for actions	0.81	0.14
3. Clarity of expectations	0.79	0.17
4. Completeness of storyline	0.78	0.28
5. Emotional responses	0.75	-0.03
6. Chronology of storyline	0.62	0.43
7. Concreteness of account	0.22	0.81
8. Consistency of evaluation	0.06	-0.59

Source: Calculated from authors' own data.

We identified three forms of *relevance* from the literature—that is, three ways in which tellers convey information sufficient for the listener to know if the teller (1) has similar expectations for health care; (2) responds to health concerns with similar actions; and (3) has similar emotional responses to clinical experiences. Because consumers reading these patient narratives about care will value these three dimensions of relevance in highly personalized ways, we did not attempt to create a composite measure.

We coded for relevance and coherence in distinct ways. The three measures of relevance were coded textually, by counting each time a respondent conveyed information about expectations, behavioral responses, or emotional responses. By contrast, coherence encompasses the complete EPN, and required codes capable of capturing whether the account conveyed in gestalt “a rich, resonant comprehension of a singular person’s situation as it unfolds in time” (Charon 2001). Following established practices in qualitative research (Reese et al. 2011), we called these “synthetic codes.” Each of the eight narrative attributes was scored on an ordinal scale (see Appendix S2 for detailed criteria) that reflected whether that attribute was absent from the narrative, partially present, or fully present. Each elicitation was assessed in terms of its coherence and relevance, relative to that of its matching intensive interview.

Representativeness as a Goal for Elicitation Performance

Our final aspirations for EPNs were informed by the need for narratives that reflect the diversity of patients’ voices and experiences (Grob and Schlesinger 2011), both to allow individual consumers to find narratives with which they can identify (see above) and to ensure that in aggregate such narratives provide a representative portrait of patients’ experiences with clinical care.

We developed two criteria for assessing the representativeness of EPNs. The first criterion is capacity to generate (as do standardized patient experience surveys) *reliable participation* from a representative cross-section of the public (Hays 2009), and thus to ensure adequate feedback from those who are less educated, more frail, or more socially isolated (Schlesinger, Mitchell, and Elbel 2002; Schlesinger 2011). Our second criterion is *equivalent quality of narrative data*, measured by capacity of the protocol to generate equally complete, balanced, and meaningful narratives from “tellers” with diverse health and socio-economic status.

In summary, we aspired to collect narratives that are (1) *complete* (providing a full picture of the experiences that matter to the patient describing them); (2) *balanced* (accurately reflecting both positive and negative aspects of the patient’s experiences, to the extent that they exist); (3) *meaningful* (conveying a story that is coherent to other patients and allows them to assess its relevance for themselves); and (4) *representative* (capturing equally high-quality reports of experiences from patients across health status and sociodemographic subgroups).

METHODS II: THE DESIGN AND TESTING OF AN ELICITATION PROTOCOL

Our elicitation protocol development process involved two rounds of data collection, with 48 respondents in each round. Round 1 tested our initial prototype elicitation protocol (see Table 2, left). Round 2 tested a version of the protocol (Table 2, right) revised to respond to shortfalls identified by applying our criteria to the first-round data (described below). Data collection in each round yielded 48 EPNs and matched interviews.

Our design used mixed method techniques (Greene, Caracelli, and Graham 1989; Wisdom et al. 2012)—qualitative to generate and code richly textured data, and quantitative to detect patterns and assess elicitation performance. To test how our elicitation protocols might perform when implemented at a large scale, we tested them in the context of CG-CAHPS, currently completed by millions of Americans each year. This required an elicitation protocol that could be kept relatively concise. Finally, because existing patient experiences surveys are typically fielded by mail, phone, and web, and because mode effects can be significant (Greene, Speizer, and Witala 2008), we also tested in both spoken (via telephone) and written formats.

Table 2: Original and Revised Elicitation Protocols

<i>Elicitation Protocols</i>	
<i>Round 1</i>	<i>Round 2</i>
Q1. When you think back over the past 12 months, what experiences come to mind?	Q1. What are the most important things that you look for in a health care provider and his or her staff?
Q2. Please tell me about anything, or anything <i>else</i> , that has gone well in your experiences over the past 12 months. Please describe these experiences as if you were telling a friend about them, so that he or she could understand how the experiences felt to you.	Q2. When you think about the things that are most important to you, how do your provider and his or her staff measure up?
Q3. What, if anything, has gone <i>better than you expected</i> in your experiences. These can be the same experiences you just described; if so, please help me understand what made these better than expected.	Q3. Now we'd like to focus on anything that has gone well in your experiences with your provider and his or her staff over the past 12 months. Please explain what happened, how it happened, and how it felt to you.
Q4. Please tell me about any experiences, or any other experiences, that you wish had gone differently over the past 12 months. Please describe these experiences as if you were telling a friend about them, so that he or she could understand how the experiences felt to you.	Q4. Next we'd like to focus on any experiences with your provider and his or her staff that you wish had gone differently over the past 12 months. Please explain what happened, how it happened, and how it felt to you.
Q5. What, if anything, has gone <i>worse than you expected</i> in your experiences with Dr. ___ and his or her staff. These can be the same experiences you just described; if so, please help me understand what made these worse than expected.	Q5. Please describe how you and your provider relate to and interact with each other.
Q6. Think back to those experiences you wish had gone differently. What, if anything, did you or your provider do about the situation and how did it turn out? For example, did you talk to your doctor about the issue, talk to your family or friends, look for another doctor, take any other form of action, or do nothing at all?	Q6. Think back to those experiences you wish had gone differently. What, if anything, did you or your provider do about the situation and how did it turn out? For example, did you talk to your doctor about the issue, talk to your family or friends, look for another doctor, take any other form of action, or do nothing at all?
Q7. Think back to those experiences that went well. What, if anything, did you do in response to these positive experiences? For example, did you talk to your doctor about them, talk to your family or friends about them, recommend this doctor to other people, take any other form of action, or do nothing at all?	Q7. Think back to those experiences that went well. What, if anything, did you do in response to these positive experiences? For example, did you talk to your doctor about them, talk to your family or friends about them, recommend this doctor to other people, take any other form of action, or do nothing at all?

Notes. Now we want to hear how you would describe, in your own words, your experiences with this provider and his or her staff—for example, a nurse or receptionist—over the past 12 months.

Study Population

Data Sources. To ensure that EPNs reflected the full range of patient experiences, study participants were recruited randomly from an existing Internet panel of over 60,000 households developed and maintained by the Gesellschaft für Konsumforschung (GfK). Past studies have found this panel representative of the American population in demographics and health status (Chang and Krosnick 2009).² The characteristics of respondents completing elicitations compared to the U.S. population are presented in Table 3.

Table 3: Characteristics of Elicitation Sample Compared to U.S. Population

<i>Characteristic</i>	<i>Elicitation Sample</i>	<i>U.S. Population</i>
Sociodemographics*		
Age		
<30	11.0%	21.5%
30–44	17.2%	25.5%
45–60	28.2%	27.1%
>60	43.6%	25.8%
Race/ethnicity		
White	78.0%	66.0%
African American	10.3%	11.6%
Latino	7.0%	15.0%
Other	4.8%	7.5%
Education		
High school or less	38.9%	42.2%
Some college	30.0%	28.9%
College graduate	31.1%	28.9%
Health status and utilization†		
Chronic health problems		
Yes	36.6%	49.8%
No	63.4%	50.2%
MD visits in previous year		
1	21.2%	31%
2–3	42.8%	43%
4–9	28.6%	22%
>9	6.6%	4%
Time with current MD		
1 year or less	23.5%	37%
2–3 years	20.1%	19%
3–4 years	16.8%	12%
5+ years	38.8%	32%

*U.S. population statistics from the Current Population Survey, 2014.

†Health care utilization statistics from the 2014 CAHPS Database.

Inclusion and Exclusion Criteria. Eligible participants were screened to ensure that they had seen a doctor in an outpatient setting at least once in the previous year. In Round 2, to test if the protocol would work equally well with consumers across categories of health status, participants were screened into three equal-sized strata: those with chronic health conditions requiring some regular supervision, those who were treated for a serious or life-threatening condition in the prior 12 months, and those who had seen a clinician in the past year but did not have chronic or serious health issues (see Appendix S3).

Study Design

Elicitation Design Principles. Our elicitation protocols were comprised of short sequences of open-ended questions (see Table 2). To encourage reporting that reflects an accurate balance of each respondent's positive and negative experiences with care, questions equally elicited what worked well and what could have been better. To facilitate coherent narratives, questions encouraged respondents to describe experiences in detail. To foster completeness, respondents were asked about their interactions with clinicians in several different ways.

Elicitation Modalities and Randomization. To test for mode effects, we split the elicitation sample in both rounds evenly between telephone and web-based modalities. We also split the sample between versions that placed the qualitative question sequence before and after the closed-ended CAHPS questions. Eligible participants in each stratum were randomized into experimental arms (mode and question placement).

Components of the Experiment. Participants completed the CAHPS survey and elicitation questions by phone or web. Phone elicitations were conducted by trained research assistants who were instructed to read the elicitation questions verbatim, and whose work was closely monitored by senior members of the study team. Intensive interviews were conducted 2–4 weeks after the elicitation by one of two experienced qualitative researchers. The intensive interview protocol used the elicitation protocol as its starting point, but incorporated specific probes for each major domain of patient experience identified from past research (López et al. 2010), including access, physician–patient interactions, quality, and coordination of care. All phone elicitations and interviews were recorded and fully transcribed.

The interviews were always conducted after the elicitation because we anticipated the priming effect of the intensive interviews on the elicitations would be significant. Interviews were scheduled several weeks after the elicitations to reduce the impact of priming from the elicitation on the interview content. The “feel” of the interview was also significantly different: interviewers established a free-flowing, conversational tone on the telephone, in contrast to the highly structured approach used in the telephone elicitation and incorporated into the format of the web-based elicitation. The order of questions used in the interview varied, following normal conversational patterns in contrast to the elicitation; the interview also included probes that were entirely absent in the elicitation. Finally, interviewers pointed out to participants at the outset of each interview that some similar question domains had been included on the earlier elicitation, and explicitly encouraged respondents to start afresh with their depiction of clinical encounters during the past year.

Data Analysis

Coding Scheme. The coding scheme we developed (see Appendix S4) was based in part on a taxonomy of experiential domains identified from the literature. Specific textual codes in each of these domains were refined during Round 1 coding following methods of grounded theory (Corbin and Strauss 1990), as we applied early versions of the scheme to interview and EPN transcripts and then constructed new codes to capture aspects of the narratives for which we did not yet have codes. Additional synthetic codes were also developed during Round 1 as we piloted the coding scheme.

Data Analysis. Transcripts contained no identifying information or indication of assigned arm or strata of the respondent. Each was coded independently by a senior member and one additional member of a four-person coding team. While coders were being trained in Round 1, the first and second author of this paper met with the senior coder weekly to discuss discrepancies, edit the coding scheme, and ensure coder reliability. Line counts and distinct mentions associated with each code in each transcript were calculated and recorded.

Coding met conventional standards of acceptable reliability (Landis and Koch 1977; Fleiss 1981). For Round 1, kappa = 0.77 for textual codes and 0.65 for synthetic codes. For Round 2, scores were 0.79 and 0.79, respectively. Comparable kappas for the intensive interviews tended to be a bit lower, reflecting the greater complexity and nuance of these narratives. Round 1

interviews yielded a kappa of 0.74 for textual codes and 0.67 for synthetic codes, with corresponding scores of 0.75 and 0.56 in Round 2.

Refining Elicitation Protocols

Each version of the protocol was pilot tested prior to fielding, using a convenience sample of diverse respondents ($N = 19$ for Round 1, $N = 20$ for Round 2), to ensure that the questions were easily readable and interpretable by respondents, whatever their level of education or experience with health care. The 39 pilot elicitations are not included in the formal analyses presented in this paper.

Analysis of Round 1 data identified several shortfalls in the performance of EPNs relative to matching interview data (see results section below). Because phone performed better than web elicitations in Round 1, for Round 2, we incorporated prompts specifically designed to encourage more narrative detail, more explicit emotional valence, and more overall structure in written narratives. To address underperformance in the domain of patient-provider communication, we added a question prompting patients to describe how they “relate to and interact with” their providers. To obtain clearer descriptions from participants regarding their expectations for clinical care, we reworked the opening sequence of questions (Table 2).

RESULTS: COMPARATIVE EFFICACY OF ELICITATION PROTOCOLS

Detailed results from our analyses are described elsewhere. Here, we highlight findings that illustrate how the metrics we created can be used to assess elicitation protocol performance relative to hour-long interviews.

To illustrate how elicitation criteria can be used to refine protocol performance, we compare three protocols: (a) a single “open-box” question inquiring about aspects of clinical care (Question 1 in Round 1), (b) the complete question sequence that complemented this first question with four additional probes intended to evoke a balance of whatever positive and negative experiences the patient encountered (full Round 1 protocol), and (c) the revised question sequence in the second round that integrated more “scaffolding” in the form of more detailed question wordings (full Round 2 protocol).

Engaging Patients in Telling Their Stories

To establish a quantifiable metric of the capacity of an elicitation protocol to successfully engage patients in providing full narratives, we first consider the total word counts induced by each protocol. The unsolicited comments about clinicians currently available on the Internet average roughly 30 to 40 words in length (Kanouse et al. 2016). By contrast, our most refined elicitation protocol (fielded in Round 2) took just 5 min to complete by Web but generated an average of 240 words. Illustrative examples of the range of EPNs generated using our final protocol are included in Appendix S4.

We quantified the comparative performance of each of the three forms of narrative elicitation described above by analyzing each version's word count as a proportion of the word count in the matched intensive interview (Table 4). Our initial protocol elicited twice as many words as the single question, which yielded responses that averaged 64 words in length. More notably, the refined version developed for Round 2 after analyzing areas of weakness in Round 1 increased relative word count by almost 50 percent above that induced by the initial protocol.

Completeness, Balance, and Meaningfulness

The 10 domains comprising our measure of completeness offer a benchmark for how patients describe their experiences with clinicians. The salience of each domain varied by health status, as reflected in the words describing each domain in the intensive interviews (Table 5). Patients who had experienced more serious illness devoted more words to coordination, while those whose interactions with clinicians involved less serious illness were more attentive to the emotional aspects of caring.

Overall, completeness scores were highest for the domains that were most salient to respondents. In the interviews, the average patient described more than eight domains. EPNs averaged four domains—slightly lower for patients with chronic health conditions (3.7) or serious health crises in the past year (3.7) than for other patients (4.2). More than 95 percent of EPNs included at least two domains.

Significant differences among protocols are evident for substantive elicitation metrics (Table 4). There was roughly a doubling of performance between the first and second round protocols in terms of the EPNs' congruence with two key aspects of interview content (communication and

Table 4: Assessing Elicitation Performance across Protocols

Protocol Used	Performance Metrics					
	Word Count [†] (Elicitation/Interview)	Completeness [‡]		Balance [§] (Positive vs. Negative)	Expressivity [¶] (Expectations)	Meaningfulness (Narrative Texture ^{††})
	Communication	Coordination				
Q1 Only	0.27	0.11	0.94	0.03	0.14	
Round 1	0.40*	0.30*	0.99	0.14	0.56*	
Full Protocol						
Round 2	0.77**	0.75**	1.06	0.44**	0.64	
Protocol						

[†]Ratio of total word count in EPN to total word count from interview.

[‡]Congruence of reporting of this domain on both EPN and interview.

[§]Ratio of positive to negative on EPN divided by the ratio of positive to negative valence on the interview.

[¶]Ratio of expectation count on EPN to expectation count on interview.

^{††}Ratio of narrative texture synthetic code from the EPN to narrative texture synthetic code from the interview.

*Difference, Round 1 versus Q1, statistically significant, $p < .05$.

**Difference, Round 2 versus Round 1, statistically significant, $p < .05$.

Source: Calculated from authors' own data.

Table 5: Salience of Domains of Patient Experience by Health Status: As Measure by Average Proportion of Words Devoted to Domain in the Intensive Interviews

	<i>Round 2: Full Sample</i>	<i>Healthy</i>	<i>Chronically Ill</i>	<i>Seriously Ill</i>
Competent	17.50%	17.00%	16.00%	20.00%
Communication	16.50%	17.30%	14.90%	17.40%
Access	15.10%	15.00%	16.60%	13.50%
Coordination	11.00%	6.20%	14.00%	14.10%
Staff	10.30%	9.60%	10.60%	10.80%
Caring	8.10%	11.00%	6.10%	6.70%
Ample time	6.40%	8.00%	5.10%	5.50%
Thorough	4.60%	4.40%	4.70%	4.70%
Shared decisions	4.10%	2.60%	6.50%	3.20%
Orientation/style	2.20%	3.00%	2.00%	1.20%

Source: Calculated from authors’ own data.

coordination) and the extent to which patients expressed their expectations for care. Differences in coherence (reported fully in other papers but illustrated here by narrative texture) increased 15 percent from Round 1 to Round 2.

One metric showed little change: the positive–negative balance in patients’ comments. Even brief protocols (Q1 alone) evoked a mix of positive and negative commentary that closely approximated the balance evident in the matched intensive interviews.

Variation across Survey Modalities. In Round 1, the telephone protocol outperformed the web-based one on every metric besides balance (see Table 6). Subsequent revisions (Round 2) eliminated differences between phone and web with respect to completeness of EPNs. The Round 2 protocol also increased both word counts and the meaningfulness of narratives (as measured by coding) for both modalities, but it did not diminish the gap between them (Table 6, columns 1, 5, and 6).

Representativeness

We assessed representativeness by comparing relative performance of the elicitation protocol among different groups of respondents. Here, we illustrate by comparing our metrics of narrative quality (1) among healthy adults, chronically ill adults, and people who had a serious health crisis within the past year,

Table 6: Assessing Elicitation Performance across Survey Modes

Survey Modality	Performance Metrics						
	Word Counts [†] (Elicitation/Interview)		Completeness [‡]		Balance [§] (Positive vs. Negative)	Meaningfulness	
	Communication	Coordination	Communication	Coordination	Expressivity [¶] (Expectation)	Expressivity [¶] (Expectation)	Narrative Texture ^{††}
Round 1 Full	0.10	0.40	0.30	0.30	0.99	0.14	0.56
Phone	0.16*	0.59*	0.37	0.37	0.99	0.20*	0.71*
Web	0.02	0.26	0.22	0.22	1.00	0.04	0.37
Round 2 Full	0.15	0.78	0.75	0.75	1.06	0.44	0.64
Phone	0.22*	0.71	0.75	0.75	1.09	0.57*	0.81*
Web	0.09	0.83	0.75	0.75	1.04	0.31	0.48

[†]Ratio of total word count in EPN to total word count from interview.

[‡]Congruence of reporting of this domain on both EPN and interview.

[§]Ratio of positive to negative on EPN divided by the ratio of positive to negative valence on the interview.

[¶]Ratio of expectation count on EPN to expectation count on interview.

^{††}Ratio of narrative texture synthetic code from the EPN to narrative texture synthetic code from the interview.

*Difference between phone and web protocol statistically significant, $p < .05$.

Source: Calculated from authors' own data.

and (2) between those who expressed more and those who expressed less during their intensive interviews.

The protocols performed equally well on almost all measures across health status strata (Table 7). Accounts were as complete and meaningful (relative to the intensive interviews) for patients with complex health problems as for those without. EPNs from patients with chronic conditions were positively biased compared to their intensive interviews. It is unclear from data collected to date why this bias with respect to balance emerges for respondents with chronic illness, but not for those who experienced health crises in the past year.

Our respondents differed substantially with respect to how much they had to say in their intensive interviews. Those at the 75th percentile of word count spoke more than twice as much (2,275 words) as those at the 25th percentile (1,015 words). Word count across interviews did not vary by health status; relatively healthy patients spoke as much on average about their experiences with clinicians (1,624 words) as those with chronic conditions (1,680 words) and those who had experienced health crises (1,625 words).

Further Refinements

In the end, the final two questions of the elicitation sequence (see Table 2) did not add greatly to the content of the narratives. Too few respondents had responded to events—positive or negative—in their health care to make this a rich source of insight. We therefore eliminated these questions.

DISCUSSION/CONCLUSION

Narratives are compelling because they speak to both our imaginations and our analytic capacities, and because they motivate as well as inform. To be useful, narratives need to be systematically generated and rigorously analyzed according to meaningful criteria. This article describes a pioneering methodology that can be used to catalyze a shift from reliance on anecdotal, often fragmentary, volunteered comments to rigorously elicited patient narratives, the quality of which can be reliably assessed. Furthermore, the improved performance from Round 1 to Round 2 protocols demonstrates that the measurement techniques we developed for this study can be effectively used to (1) reveal areas of strength and weakness in elicitation instruments, and (2) gauge whether refinement improves them. This capacity to learn and improve is an

Table 7: Representativeness: Elicitation Performance across Illness Strata

Strata of Health Needs	Performance Metrics						
	Word Counts †		Completeness ‡		Balance §	Meaningfulness	
	(Elicitation/Interview)	Communication	Coordination	(Positive vs. Negative)	Expressivity¶ (Expectations)	Narrative Texture ††	
Healthy	0.17	0.81	0.56*	1.02	0.37	0.66	
Chronically ill	0.14	0.76	0.82	1.17	0.37	0.52	
Seriously ill	0.13	0.73	0.87	1.00	0.61	0.76	

†Ratio of total word count in EPN to total word count from interview.

‡Congruence of reporting of this domain on both EPN and interview.

§Ratio of positive valence over negative valence on EPN, divided by the comparable ratio for the interview.

¶Ratio of expectation count on EPN to expectation count on interview.

††Ratio of narrative texture synthetic code from the EPN to narrative texture synthetic code from the interview.

*Difference between more and less seriously ill statistically significant, $p < .05$.

Source: Calculated from authors' own data.

essential building block for the move from anecdote to science in the elicitation of patient narratives.

Previous articles have analyzed the content of anecdotal patient commentary (López et al. 2010; Gao et al. 2012; Detz, López, and Sarkar 2013; Greaves, Millett, and Nuki 2014). To our knowledge, this is the first effort to go beyond assessment to foster narratives that are more complete, meaningful, balanced, and representative. We define criteria for assessing our success, based on both the content and the narrative characteristics of patient accounts.

Our findings here provide a promising base on which additional research about methods for and implications of publicly reporting narrative data can be built. Earlier research has demonstrated the power of narrative data to enhance public engagement with quality reporting websites (Schlesinger et al. 2014); what will be needed now is research about how best to present narrative data, and what impact such data have on consumer decision making (Schlesinger, Grob, and Shaller 2015b).

These findings should be interpreted in light of several limitations. First, our gold standard comparing EPNs with patient interviews reflects the “truth” of clinician interactions as the patient understands them, but these descriptions might differ from those provided by a clinician or other third-party observer. Second, our taxonomy of domains assessing completeness would not necessarily apply in other clinical settings, such as hospitals or long-term care. Third, using a standing Internet panel for collecting data allows comparisons of subgroups of respondents in terms of how complete, meaningful, and balanced their elicited accounts are, but not of participation rates (as participant responses are incentivized). Fourth, CAHPS and similar patient experience surveys are still administered predominantly by mail, and our study did not test a mail modality. Finally, by assessing the communicative capacity of EPNs as a proportion of what can be discerned from the matching interview, we can assess the “goodness” of elicitation protocols in a relative sense, but not establish what degree of congruence is “good enough” to merit the time, effort, and resources required to collect them.

The work we report here is intended as a foundation for a broader agenda of future research and field testing. We analyze differences in elicitation performance by sociodemographic status in other papers, where we demonstrate that there are no substantial differences by age, educational attainment, or gender.

The Round 2 elicitation protocol presented above is best understood as a prototype that can and should be further refined, as well as adapted for use in other clinical settings. Substantial ongoing investment in the continued

development of elicitation methods will be necessary, especially as additional insights are gleaned from research in the field. As noted above, we also have more to learn about effectively reporting EPNs in conjunction with other performance metrics (Schlesinger et al. 2014, 2015a). Given the potential of narratives to enrich public reporting, better inform consumer choices, facilitate quality improvement, and improve our understanding of patients' expectations, experience, and well-being, we believe such an investment is well merited.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: The research reported here benefited from the contributions of Karin Liu as an interviewer as well as the research assistance of Erika Rogan, Chaz Felix, and Stephanie Huang. This paper was supported by a grant to Yale University from AHRQ (1R21HS021858) and two cooperative agreements (2U18HS016980 and 1U18HS016978) from AHRQ to RAND and Yale University, respectively.

Disclosures: None.

Disclaimers: None.

NOTES

1. <http://www.yourdictionary.com/anecdotal>
2. The panel is constructed using address-based sampling to represent the noninstitutionalized U.S. population, including listed and unlisted phone numbers, cell-phone-only households, and households without phone lines. Households without access to computers are given access through computers supplied by GfK.

REFERENCES

- Chang, L., and J. A. Krosnick. 2009. "National Surveys Via RDD Telephone Interviewing versus the Internet: Comparing Sample Representativeness and Response Quality." *Public Opinion Quarterly* 73: 641–78.
- Charon, R. 2001. "Narrative Medicine: A Model for Empathy, Reflection, Profession, and Trust." *Journal of the American Medical Association* 286 (15): 1897–902.
- Cleary, P. D., C. Crofton, R. D. Hays, and R. Horner. 2012. "Advances from the Consumer Assessment of Healthcare Providers and Systems Project." *Medical Care* 50 (Suppl): S1.

- Cognetta-Rieke, C., and S. Guney. 2014. "Analytical Insights from Patient Narratives: The Next Step for Better Patient Experience." *Journal of Patient Experience* 1: 22–4.
- Corbin, J., and A. Strauss. 1990. "Grounded Theory Research: Procedures, Canons and Evaluative Criteria." *Qualitative Sociology* 13: 3–21.
- Detz, A., A. López, and U. Sarkar. 2013. "Long-Term Doctor-Patient Relationships: Patient Perspective from Online Reviews." *Journal of Medical Internet Research* 15 (7): e131.
- Dieckmann, N. F., P. Slovic, and E. M. Peters. 2011. "The Use of Narrative Evidence and Explicit Likelihood by Decisionmakers Varying in Numeracy." *Risk Analysis* 29: 1473–88.
- Fleiss, J. L. 1981. *Statistical Methods for Rates and Proportions*, 2nd Edition. New York: John Wiley.
- Galizzi, M. M., M. Miraldo, C. Stavropoulou, M. Desai, W. Jayatunga, M. Joshi, and S. Parikh. 2012. "Who Is More Likely to Use Doctor-Rating Websites, and Why? A Cross-Sectional Study in London." *BMJ Open* 2 (6): e001493.
- Gann, B. 2013. "Understanding and Using Health Experiences: The Policy Landscape." In *Understanding and Using Health Experiences*, edited by S. Ziebland, A. Coulter, J. D. Calabrese, and L. Locock, pp. 150–61. Oxford, UK: Oxford University Press.
- Gao, G. G., J. S. McCullough, R. Agarwal, and A. K. Jha. 2012. "A Changing Landscape of Physician Quality Reporting: Analysis of Patients' Online Ratings of Their Physicians over a 5-Year Period." *Journal of Medical Internet Research* 14 (1): e38.
- Greaves, F., C. Millett, and P. Nuki. 2014. "England's Experience Incorporating "Anecdotal" Reports from Consumers into Their National Reporting System: Lessons for the United States of What to Do or Not to Do?" *Medical Care Research and Review* 71 (5): 65S–80S.
- Greene, J. C., V. J. Caracelli, and W. F. Graham. 1989. "Toward a Conceptual Framework for Mixed-Method Evaluation." *Educational Evaluation and Policy* 11: 255–74.
- Greene, J., H. Speizer, and W. Wiitala. 2008. "Telephone and Web: Mixed-Mode Challenge." *Health Services Research* 43: 230–48.
- Greenhalgh, T., and B. Hurwitz. 1999. "Narrative Based Medicine: Why Study Narrative?" *British Medical Journal* 318 (7175): 48.
- Grob, R., and M. Schlesinger. 2011. "Epilogue: Principles for Engaging Patients in U.S. Health Care and Policy." In *Patients as Policy Actors: A Century of Changing Markets and Missions*, edited by B. Hoffman, N. Tomes, R. Grob, and M. Schlesinger, pp. 278–91. New Brunswick, NJ: Rutgers University Press.
- Hays, R. 2009. "Patient Satisfaction." In *Encyclopedia of Medical Decision Making*, edited by M. W. Kattan, and M. E. Cowen, pp. 866–8. Thousand Oaks, CA: Sage.
- Health Research Institute. 2013. *Scoring Healthcare: Navigating Customer Experience Ratings*. Delaware: Pricewaterhouse Coopers.
- Kaiser Family Foundation. 2015. *Kaiser Health Tracking Poll: April 2015*. Publication No. 8718-T. Menlo Park, CA: Kaiser Family Foundation.

- Kanouse, D. E., M. Schlesinger, D. Shaller, S. C. Martino, and L. Rybowski. 2016. "How Patient Comments Affect Consumers' Use of Physician Performance Measures." *Medical Care* 54(1): 24–31.
- Lagu, T., and P. K. Lindenauer. 2010. "Putting the Public Back in Public Reporting of Health Care Quality." *Journal of the American Medical Association* 304: 1711–2.
- Lagu, T., S. L. Goff, N. S. Hannon, A. Shatz, and P. K. Lindenauer. 2013. "A Mixed-Methods Analysis of Patient Reviews of Hospital Care in England: Implications for Public Reporting of Health Care Quality Data in the United States." *Joint Commission Journal on Quality and Patient Safety* 39 (1): 7–15.
- Landis, J. R., and G. G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33 (1): 159–74. doi:10.2307/2529310.
- Lansky, D. 2012. "Public Reporting of Health Care Quality: Principles for Moving Forward." *Health Affairs* [accessed on February 2, 2015]. Available at <http://healthaffairs.org/blog/2012/04/09/public-reporting-of-health-care-quality-principles-for-moving-forward/>
- López, A., A. Detz, N. Ratanawongsa, and U. Sarkar. 2010. "What Patients Say about Their Doctors Online: A Qualitative Content Analysis." *Journal of General Internal Medicine* 27: 685–92.
- Marshall, M., and V. McLaughlin. 2010. "How Do Patients Use Information on Health Providers?" *British Medical Journal* 341: c5272.
- McAdams, D. P. 2006. "The Problem of Narrative Coherence." *Journal of Constructivist Psychology* 19 (2): 109–25.
- Mitchell, S., and M. Schlesinger. 2005. "Managed Care and Gender Disparities in Problematic Healthcare Experiences." *Health Services Research* 40 (5): 1489–513.
- Reese, E., C. A. Haden, L. Baker-Ward, P. Bauer, R. Fivush, and P. A. Ornstein. 2011. "Coherence of Personal Narratives across the Lifespan: A Multidimensional Model and Coding Method." *Journal of Cognition and Development* 12 (4): 424–62.
- Schlesinger, M. 2011. "The Canary in the Gemeinschaft: The Public Voice of Patients as a Means of Enhancing Health System Performance." In *Patients as Policy Actors*, edited by B. Hoffman, N. Tomes, R. Grob, and M. Schlesinger, pp. 148–76. New Brunswick, NJ: Rutgers University Press.
- Schlesinger, M., R. Grob, and D. Shaller. 2015b. "Using Patient Reported Information to Improve Clinical Practice." *Health Services Research* 50 (S2 Pt II): 2116–54.
- Schlesinger, M., S. Mitchell, and B. Elbel. 2002. "Voices Unheard: Barriers to Expressing Dissatisfaction to Health Plans." *Milbank Quarterly* 80 (4): 709–55.
- Schlesinger, M., D. Kanouse, S. Martino, L. Rybowski, and D. Shaller. 2014. "The Effects of Complexity on Consumers' Choice of Doctors: A Look inside the Blackest Box." *Medical Care Research & Review* 71 (5): 38S–64S.
- Schlesinger, M., R. Grob, D. Shaller, S. C. Martino, A. M. Parker, M. L. Finucane, J. L. Cerully, and L. Rybowski. 2015a. "Taking Patients' Narratives about Clinicians from Anecdote to Science." *New England Journal of Medicine* 373 (7): 675–9.

- Shaffer, V. A., S. Tomek, and L. Hulsey. 2014. "The Effect of Narrative Information in a Publicly Available Patient Decision Aid for Early-Stage Breast Cancer." *Health Communication* 29 (1): 64–73.
- Shaffer, V. A., and B. J. Zikmund-Fisher. 2013. "All Stories Are Not Alike: A Purpose-, Content-, and Valence-Based Taxonomy of Patient Narratives in Decision Aids." *Medical Decision Making: An International Journal of the Society for Medical Decision Making* 33 (1): 4–13.
- Sick, B., and J. M. Abraham. 2011. "Seek and Ye Shall Find: Consumer Search for Objective Health Care Cost and Quality Information." *American Journal of Medical Quality* 26: 433–40.
- Tompson, J., and N. Wilcoxon. 2014. *Finding Quality Doctors: How Americans Evaluate Provider Quality in the United States*. Chicago, IL: NORC [accessed on November 11, 2014]. Available at <http://www.apnorc.org/projects/Pages/finding-quality-doctors-how-americans-evaluate-provider-quality-in-the-united-states.aspx>
- Trigg, L. 2011. "Patients' Opinions of Health Care Providers for Supporting Choice and Quality Improvement." *Journal of Health Services Research and Policy* 16: 102–7.
- Tsianakas, V., J. Maben, T. Wiseman, G. Robert, A. Richardson, P. Madden, M. Griffin, and E. A. Davies. 2012. "Using Patients' Experiences to Identify Priorities for Quality Improvement in Breast Cancer Care: Patient Narratives, Surveys or Both?" *BMC Health Services Research* 12 (1): 271.
- Tu, H., and J. May. 2007. "Self-Pay Markets in Health Care: Consumer Nirvana or Caveat Emptor?" *Health Affairs* 26 (2): w217–26.
- Vasquez, C. 2014. *The Discourse of Online Consumer Reviews*. London: Bloomsbury.
- Winterbottom, A., H. L. Bekker, M. Conner, and A. Mooney. 2008. "Does Narrative Information Bias Individual's Decision Making? A Systematic Review." *Social Science & Medicine* 67 (12): 2079–88.
- Wisdom, J. P., M. A. Cavaleri, A. J. Onwuegbuzie, and C. A. Green. 2012. "Methodological Reporting in Qualitative, Quantitative and Mixed Methods Health Services Research Articles." *Health Services Research* 47: 721–44.
- Ziebland, S., and S. Wyke. 2012. "Health and Illness in a Connected World: How Might Sharing Experiences on the Internet Affect People's Health?" *Milbank Quarterly* 90 (2): 219–49.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Appendix S1: Ten Domains of Clinician–Patient Interaction.

Appendix S2: Eight Dimensions and Coding for Synthetic Codes on Narratives.

Appendix S3: Screening Questions for Stratifying Round 2 Elicitation Sample.

Appendix S4: Complete Coding Scheme: Textual Codes.

Appendix S5: Illustrative Examples of Elicited Narratives.