Can Planned Visit Scores Predict Practice Improvements in Patient Experience Outcomes? A Meta-Regression Study

Gerald K Arnold[1], Rebecca A Baranowski[1], and Lauren Duhigg[1]
[1]American Board of Internal Medicine, 510 Walnut Street, Suite 1700, Philadelphia, PA 19106-3699

Abstract
For patients with chronic medical conditions, the goal of planned visits is to ensure that clinical teams are prepared for the visit, so that patients receive evidence-based care and condition-specific self-management training. Few studies on the effectiveness of planned visits exist; instead, most researchers have focused on the overall Chronic Care Model (Coleman, et al. 2009), of which planned visits are only one part.

We (Arnold, Baranowski, and Duhigg, 2013) investigated whether primary care practices utilizing all components of a planned visit protocol (PVP) received better patient-experience scores than matched-control practices not using a PVP. Better scores included higher ratings on quality of physician communication, shared decision-making, interactions with practice staff, and overall quality of care. Our findings showed that patients in practices utilizing a PVP gave significantly higher ratings on two measures: physicians involving patients in their care plan (shared decision-making, p=.005) and office staff being helpful and courteous (interactions with staff, p<.001).

The current study included 882 internists from practices reporting specific chronic illnesses (such as diabetes or hypertension) as the most important condition among their patients. These physicians completed the American Board of Internal Medicine's (ABIM) Practice Improvement Module (PIM)® on physician-patient communication in primary care, as part of the Maintenance of Certification program. Eligible patients, who made three or more office visits per year, completed CG-CAHPS®–based surveys. There were 15,318 patient responses at baseline and 14,146 at follow-up after practice-improvement efforts. A PVP reportedly was used by 128 physicians; 754 practices without PVP served as controls. A second control group included 128 practices propensity matched to the PVP cases to control for practice, physician, and patient differences. We used five patient-experience composite measures: 1) physician communication, score range 6-36, Cronbach's α = .87; 2) level of shared decision-making (score range 1-8, Cronbach's α = .65; 3) helpfulness/courteousness of staff (score range 1-12, Cronbach's α = .84); 4) patient follow-up on lab results score range1-6, Cronbach's α = .86); and 5) overall physician rating score, range 0-10, Cronbach's α = .89). We hypothesized that practices using a PVP would have different trends in Spearman correlation effect sizes from their quality-improvement exercises than practices not using a PVP. These effect size trends could help estimate minimally important differences for patient-experience outcome measures, identify procedures that practices could use to improve their patient experience scores, and test whether summary measures could reproduce effect size estimates similar to those obtained from mixed linear models.

Spearman correlation effect size estimates for each practice, converted to Fisher Z estimates, served as the dependent variable. Weighted Fisher Z estimates were combined to create an overall effect size estimate for patient experience along with separate Z estimates for each composite. The independent variable was an indicator variable designating if a practice followed a PVP. Since each estimate represented an independent rapid-cycle test of change at the practice level, analysis followed a meta-regression procedure fitting both fixed effects regression and random effects via Der Simonian & Laird (1986) procedures.

Random effects models fit the data better than the fixed effects models, showing that effect sizes were sufficiently homogeneous to compare across practices. The PVP practices had significantly better effect sizes for overall patient experience (p=.005) and shared decision-making (p=.001) but not for physician communication (p=.422), helpfulness/courteousness of staff (p=.473), follow-up on lab results (p=.930), and overall physician rating (p=.119). After adjusting for physician, practice, and patient differences through matching the significance of the overall patient experience, effect size was reduced (p=.069), but the significant effect size for shared decision-making remained (p=.001).

Practices following a PVP show significantly greater effect size gains in overall patient experience and shared decision-making. For minimally important differences in patient experience, effect sizes are about .03 Fisher Z units or .067 standard deviation units (Cohen's D). Effect size measures based on summary data such Spearman correlations can approximate results from random effects models that use patient level responses. This study suggests that PVPs are a specific set of procedures that primary care practices can implement to improve their patient-experience scores.

Key Words: Spearman correlation, effect size, meta-regression, propensity scores

1. Introduction

The Chronic Care Model (Wagner, 2007) plays a dominant role in assessing health care in this country. Research on the Chronic Care Model (CCM) has typically focused on the model as a whole (Coleman, et al. 2009).

One aspect of CCM is the concept of planned visits (Edgeman-Levitan, et al., 2003 & 2012). This means that physicians and non-physician providers work together as a team to provide care that is evidence-based and focused on motivating patients to be involved in their own treatment plans and self-care. The planned visit protocol (PVP) enables the patient in a patient-centered manner. Generally, planned visits carried out at the practice level include the following components:

- Visits are arranged and conducted by practice teams
- Team members work with patients to set and monitor care goals
- Care givers provide evidence-based care
- Teams provide condition-specific self-management training tailored to each patient

Some researchers have predicted that a PVP should improve patient satisfaction and their overall experience with care (Glaseroff, 2007). We are interested in how planned visits affect individual practice performance with regard to patient experience measures. Our team (Arnold, Baranowski & Duhigg, 2103) previously demonstrated that practices using planned visits had higher baseline patient experience measures than practices that did not follow a PVP, after accounting for differences in physician and practice characteristics and adjusting for variation in patient case-mix among practices. In this study we investigate whether medical practices that report using a PVP have associated improvements in patient-experience measures at the time of re-measurement, after participation in quality-improvement activities.

1.1 Effect Sizes and Hypotheses Related to Planned Visits in Practices

Changes in patient-experience measures from baseline to follow-up are converted to correlational effect sizes (Rosenthal, Rosnow, and Rubin, 2000). We examined whether primary care practices that report using a PVP have effect sizes related to patient-experience measures that differ from practices not following PVP. We expected measures of patient experience related to provider communication, shared decision-making, staff being respectful, and overall rating of the provider to be influenced by a PVP. We did not expect patients' experience with receiving lab results to be affected by a PVP. This gives us an opportunity to investigate if whether differential results can be observed from planned visits, further validating the protocol. The importance of this study is the demonstration that a specific aspect of the Chronic Care Model (Wagner, 2007) implemented at the practice level, can have a direct effect on improving the patients' reports of their experiences with the care they receive.

There are two advantages of using effect sizes (Z) in PVP comparisons. First, individual measures of patient experience may be examined, and these measures can be combined into a single, overall measure of the patients' reported experiences of care. Our hypotheses in this study were:

Ha: Overall patient experience; $Z_{pv} \neq Z_{control}$
Ha: Provider (physician) communication skills; $Z_{pv} \neq Z_{control}$
Ha: Physician involves patients in treatment decisions (shared decision-making); $Z_{pv} \neq Z_{control}$
Ha: Staff were respectful and helpful to patients; $Z_{pv} \neq Z_{control}$
Ha: Patients were provided test results in follow-up; $Z_{pv} \neq Z_{control}$
Ha: Overall rating of physician; $Z_{pv} \neq Z_{control}$

Where:
Z_{pv} = effect size for practices using PVPs
$Z_{control}$ = effect size for practices not using PVPs

The second advantage of using effect sizes is that regardless of the quality improvement activities of individual practices, outcomes from those activities can be put on the same scale. Thus, this study is similar to a meta-analytic study. Because we are using weighted regression as our principal analytic tool, we called the study a meta-regression.

1.2 Context and Study Setting

Our study examines the American Board of Internal Medicine (ABIM) Maintenance of Certification program. Board certification at ABIM is time limited; physicians in internal medicine and its subspecialties must retake a secure exam every ten years. They must maintain a valid medical license and must demonstrate on a periodic basis that they are actively engaged in self-evaluation of their medical knowledge and quality of care, called practice assessment. One of our tools for practice assessment is the Practice Improvement Module (PIM)®.

In this study we focused on the assessment of a physician's communication skills with patients, using the Communication PIM for Primary Care, shown in Figure 1. Our patient-experience measures come from the Clinician and Group Consumer Assessment of Healthcare Providers and Systems (CG-CAHPS®) survey. The survey item pool is endorsed by the National Quality Forum and is used for patent-experience quality measures in large systems such as Medicare.

CG-CAHPS® are administered by the practices. Patients are typically directed to a website for administration or call a toll-free number to have the instrument administered via telephone.

The baseline data collection phase of the Communication PIM includes a very detailed questionnaire about practice operations, interactions with patients, practice infrastructure, and the nature of a practice's quality-improvement activities. This systems survey is based on the National Committee for Quality Assurance ((NCQA 2014)) Physician Practice Connections® (PPC®) program (see http://www.ncqa.org/Portals/0/Programs/Recognition/PPC_web.pdf). We derived the planned visit score from this questionnaire.

An initial sample of at least 25 patients is collected for baseline measurement of patient experience. ABIM receives this information over the web, compiles it, and presents it back to the physician.

The physician picks a single item from the patient questionnaire to improve, sets a goal, and develops an intervention plan. That plan is submitted on the web. On average, the implementation period is about three months. Then, a second sample of 25 patients is administered CAHPS. The data are compiled and given back to the physician, who considers the results, reports them, and reflects on the success of the improvement process. The PIM is based on a standard Shewhart-Deming rapid-cycle process of measurement,, intervention, re-measurement, and review (ASQ 2014 see http://asq.org/learn-about-quality/project-planning-tools/overview/pdca-cycle.html). As such, each practice conducts its own quality-improvement study. Because of the PIM structure we believed the methods of meta-analysis, particularly meta-regression, could be applied to these data (Sutton, et al. 2000).

As a general rule, physicians almost always improve on the particular survey item they select. This is a self-evaluation process, not a scored summative one. We were not interested in the item they improved but rather whether practices that used a PVP showed improvement on composites of patient-experience measures regardless of whether the item selected for the PIM exercise was included in a composite.

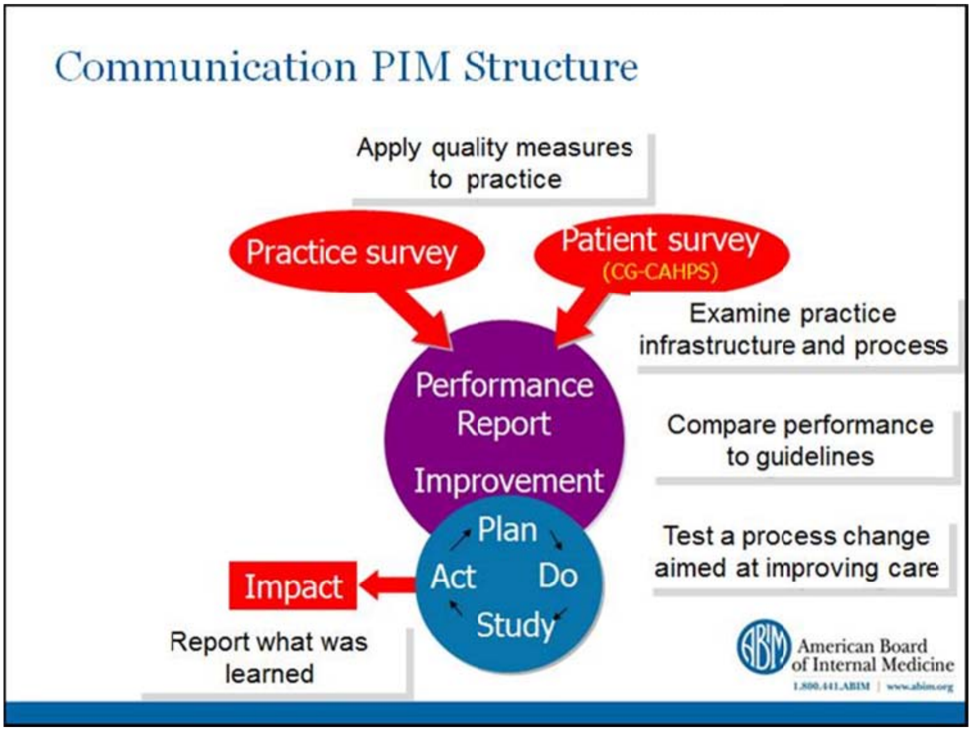

Figure 1: Structure and Steps in the Communications Practice Improvement Module

2. Data and Methods

2.1 Measures of Planned Visits and Patient Experience of Care

The primary measures used in this study were the PVP scale derived from the detailed PPC-based, systems survey mentioned earlier and five composite measures of patient experience taken from the CG-CAHPS®, Adult Primary Care Questionnaire 1.0, English language version with 6-point response scales (AHRQ, 2012). One section of the systems survey, care management and patient self-care, queries practice staff about issues related to planned visits.

2.1.1 Planned Visit Protocol Scale:

Table 1 lists the 21 items and weights used to create the PVP scores. Respondents to the survey indicated whether the practice carried out the tasks specific to care management and patient self-care. There were ten required items, which reflected the care characteristics of practices following a PVP described in "The CAHPS® Improvement Guide Practical Strategies for Improving the Patient Care Experience" (Edgeman-Levitan, et al., 2003 and 2012).

We weighted the 21 items so that the PVP score was scaled from 1 to 100. Item weights were determined by regressing the normalized ranks (Blom ranks) on the sums of the 21 items scored with a 0 (absent) or 1 (present) format. The standardized regression weights were rescaled so that their sum would equal 100 if all items were answered positively. We defined practices that engaged in PVPs as those whose scores were 75 and higher and who indicated that they carried out the 10 required items. Practices with lower scores or those with fewer than the 10 required items were categorized as not having PVPs.

Table 1. Items with (Weights) Related to Planned Visit Protocol for Patients with the Most Import Conditions Treated by the Practice	
Required Items (46 points *)	Supplemental Items (53 points *)
1. Assure all needed information for patient visit is available at time of visit (5)	1. Review longitudinal data on targeted clinical measurements for patients (3)
2.Review and individualize care management plans for patients (3)	2. Use registries to identify important medical conditions presenting to the practice (5)
3. Help patients set individualized treatment goals (6)	3. Use evidence-based guidelines for care (5)
4. Assess and document patients progress toward treatment goals (5)	4. Use templates, checklists, and guides for treating important conditions (7)
5. Review all prescribed medications, supplements, and alternative treatments with pts. at each visit (8)	5. Provide patients with instruction in self-management skills (5)
6. Review self-monitoring results for pts. document in records (3)	6. Provide pts. with self-monitoring records (5)
7. Assess barriers when patients have not met treatment goals (4)	7. Connect patients to support groups related to their conditions (4)
8. Assess barriers when patients have not filled, refilled or taken prescribed medications (3)	8.Provide access to qualified instructors for self-care (5)
9. Follow-up with patients that have not keep important appointments (5)	9. Connect pats. to self-management resources (3)
10. Conduct after visit follow-up with pts. (4)	10. Assess patients' readiness to change for participation in self-care behaviors (4)
	11. Provide care information in the patients' native languages (7)

*Note: Sum of item weights = 99; 1 is added to all scores so the range is between 1 and 100.

2.1.1 Patient-Experience Composite Measures and Effect Sizes:

We chose five composite measures to represent the patient experience in our study practices: 1) physician communication, 6 items on 6-point scale (never to always), score range 6-36, Cronbach's $\alpha = .87$; 2) level of shared decision-making, 1 (yes/no) item and 2 4-point items (definitely yes to definitely no), score range 0-9, Cronbach's $\alpha = .65$; 3) helpfulness/courteousness of staff, 2 6-point items (never to always), score range 1-12, Cronbach's $\alpha = .84$; 4) follow-up on lab results 1 (never to always) item, score range1-6, Cronbach's $\alpha = .86$; and 5) overall physician rating, score range 0-10; 0 representing the worst doctor and 10 the best, Cronbach's $\alpha = .89$. All composite reliabilities are derived from patient samples collected at the ABIM.

We coded the measurement at baseline as time period zero and re-measurement as time period one, and calculated Spearman correlation effect sizes between the ranks of the composite scores and the time measure. We used Spearman because the composite scores were considered at least ordinal. The correlation effect sizes can be transformed into Fischer Z statistics, which have an approximate normal distribution with variance estimates (Sheskin, 2004). We used the Zar (1999) correction to the variance for the Spearman correlation to account for the inefficiency due to the use of ranks (i.e., variance Z = 1.06/(n-3)) . Fischer Z can easily be converted back into a Spearman correlation, as well as other effect size measures such as Cohen's D, since the biserial correlation is in fact equivalent to a t-test using the rank transformations to the data.

With a variance estimate for each composite we can compute an overall, weighted average patient experience effect size across the five composites for each physician.

2.1.2. Study Design, Populations, and Samples

The study could be described as having a Pretest-post-test, control group, nested (patients within physicians) cross-sectional design (Murray, 1998). From a pool of about 1,400 primary care internists who completed the Primary Care Communication PIM, we picked 882 who completed the module between April 2010 and November 2013. Eligible patients made three or more office visits per year. There were 29,464 eligible patient questionnaires (pre and post); 15,318 baseline responses and 14,146 responses at follow-up after practice-improvement efforts. All 882 practices received a planned visit score and 128 of these practices met criteria for PVPs. The remaining 754 were used as controls.

We reduced the number to 882 because the PIM was changed in April 2010. Previously, there were no re-measurement samples. Physicians merely picked a single item on which to improve performance and reported whether they had met their goal or not. In 2010, we revised the PIM so complete samples would be picked for both observation periods. This enabled us to calculate the score gain effect sizes on entire composites.

Effect sizes can differ because physicians vary with respect to training, practice settings, and demographics. Likewise patients seen in these practices vary by demographics, health status, and education. To correct for these discrepancies among practices we selected 128 matched-control practices using a shortest Mahalanobis distances(Rosenbaum and Rubin 2006 & Austin, 2010), greedy match program developed by Kosanke and Bergstrath at the Mayo Clinic (2003).

We calculated dual-propensity scores correcting for 31 physician and practice characteristics and 11 patient characteristics. Dual-propensity score matching formed the case control pairs of practices. The physician-level propensity scores, calculated by logistic regression, included 31 test performance, practice characteristic, demographic, and PIM version variables. The second score included nine patient demographic and PIM version variables and all two-way interactions among these variables. Patients are nested within physicians, so the physician level score was a weighted average of patient responses adjusted for within-physician clustering. Practice pairs were matched between 128 case and 128 control practices.

We dummy coded the PV practices and used a weighted least-squares regression via SURVEYREG in SAS to predict Fisher Z's for each composite and the overall patient experience effect size. We used a fixed-effect regression and repeated the analysis with a random-effects regression using the DerSimonian and Laird method (1986). We confirmed appropriateness of the models used. We assessed model fit and effect-size homogeneity (fixed versus random) using Cochran's Q (Cochran, 1954) .

3. Results

3.1 Homogeneity of Effect Sizes and Population Effect Size Estimates

Table 2 shows the results from Cochran's Q heterogeneity Chi-squared statistics. The random-effects model consistently fits the data better than the fixed-effects model. The Cochran Q statistic indicates that effect sizes for these studies are sufficiently homogeneous to make comparisons among practices using the random-effects model.

Table 2: Homogeneity Tests and Effect Sizes Estimates for Pre- and Post-Test Changes in Patient-Experience Composites

Composite Domain	***Fixed Effects***	***Random Effects***	***Composite Effect Sizes***			
	$X^2_{(df)}$, prob.	***$X^2_{(df)}$, prob.***	***Fischer's Z***	***Z 95% CI***	***Spearman Correlation***	***Cohen's D***
Overall Experience	$\chi^2_{(4255)}$ = 11,288.36, p<.001	$\chi^2_{(4255)}$ =4189.86, p=.759	0.033	.024-043	.033	.067
Provider Communication	$\chi^2_{(867)}$ = 2,965.97, p<.001	$\chi^2_{(867)}$ = 856.50, p=.594	0.026	.003-.049	.026	.052
SharedDecision-making	$\chi^2_{(854)}$ = 1,785.19, p<.001	$\chi^2_{(854)}$ = 830.96, p=.708	0.034	.015-.054	.034	.069
Staff Respectful & Helpful	$\chi^2_{(851)}$ = 2,470.37, p<.001	$\chi^2_{(851)}$ = 829.32, p=.696	0.042	.021-.063	.042	.084
Follow-up Labs	$\chi^2_{(813)}$ = 1,999.85, p<.001	$\chi^2_{(813)}$ = 800.18, p=.619	0.058	.037-.079	.058	.112
Doctor Rating	$\chi^2_{(866)}$ = 2,043.26, p<.001	$\chi^2_{(866)}$ = 861.30, p=.539	0.011	-.009-.030	.011	.021

The Fisher Z statistics in Table 2 are the effect sizes for all 882 practices. The effect size for Overall Experience is the weighted average for the five composites that follow. With the exception of the 11-point doctor rating, all effect sizes are positive. The table shows that the simple process of re-measuring is enough to raise effect sizes for composites across all domains. In this study we want to see if or how much of an effect size increase there is among practices with PVPs. However, these figures suggest a minimally important difference for measuring changes in patient experience composites. Cohen's D is the standardized mean difference between the pre- and post-test; these estimates can be considered minimally important differences. On average (Overall Experience) the meaningful differences between means should be at least seven one-hundredths (.067) of a standard deviation.

Table 3 shows the results of the weighted ANOVA comparisons between the practices that report following a PVP and practices that do not. The rows labeled "Population" show differences in effect sizes between the 128 PVP practices and the remaining 754 control practices. If the effect sizes are positive, the increase favors the PVP practices; negative values favor the controls. These controls are not adjusted for physician and patient characteristics. The rows labeled "Matched Controls" are the differences between the PVP practices and the 128 matched control practices adjusted for physician and patient differences.

We see that without adjustment PVPs have higher gains in overall experience, communication, shared decision-making, staff being helpful, and overall physician rating. Reporting results of lab tests favored controls, but the result is near zero.

After adjustment, the overall experience gain is maintained along with shared decision-making and overall physician rating. Shared decision-making favors PVP practices significantly after adjustment for physician and patient factors. Overall experience also favors PVP practices, but the significance is more sensitive to practice factors and perhaps other biases.

Table 3: Weighted Regression Results Comparing PVP Practices with Controls: β Indicates Gain in Effect Size (+) of PVP Practices over controls where (-) Gain of Controls over PV Practices

Composite Domain	***128 Planned Visit Practices Versus:***	***Regression Coefficients For Planned Visits (PV = 1 versus Controls = 0)***		
		β	***SE(β)***	***P-value***
Overall Experience	Population (N=754)	.042	.015	.005**
	Matched Controls (n=128)	.034	.019	.069*
Provider Communication	Population (N=754)	.032	0.04	0.422
	Matched Controls (128)	-.002	.048	.963
SharedDecision-making	Population (N=754)	.097	.029	.001**
	Matched Controls (128)	.124	.037	.001**
Staff Respectful & Helpful	Population (N=754)	.028	.039	.473

	Matched Controls (128)	-.015	.046	.749
Follow-up Labs	Population (N=754)	-.003	0.031	.930
	Matched Controls (128)	-.015	0.039	.709
Doctor Rating	Population (N=754)	.046	0.03	.119
	Matched Controls (128)	.052	.038	.171

4. Discussion and Conclusions

Practices utilizing a PVP tend to boost the patient-experience effect sizes following a quality-improvement intervention at the practice level, especially for ratings of physicians involving their patients in health care decisions. The gains in the shared decision-making composites were robust with respect to patient and physician practice differences. Other scores were sensitive to these factors. As expected, we did not see an improvement in patients' ratings regarding receiving lab test results because PVPs do not directly address this. Our hypothesis tests regarding provider communication, helpful and courteous staff, and overall doctor rating were inconclusive.

There are limitations to this observational study. These studies are all based on self-reported data. The results are not audited so they may have inaccuracies. Physicians who participated in the Communication PIM are a self-selected sample of primary care practices so it is unknown to what extent these results generalize to other practice or care settings. Many patient-experience measures were sensitive to practice and patient characteristics. They may be sensitive to other factors that we did not consider.

Our future research includes comparing our summary, "random-effects" adjustment results to other models such a restricted maximum likelihood used in mixed-model assessments. Can we reproduce these results using other procedures? If so, then analyses using summary statistics from practices may be possible. We also want to see if these PVP results persist over time and whether they apply to subspecialty practices focusing on specific patient populations, such oncology or cardiology. Finally we would like to compare the effect sizes achieved from PVPs with other practice-level efforts to improve the patient experience. For example, how do PVPs compare with open-access programs of care (Edgeman-Levitan, et al., 2003 and 2012)?

The PVP appears to have a positive effect on patient-experience scores. The effect a PVP has on patients may also have a positive feedback effect on practice staff. If patients follow treatment plans and believe they play an important role in their own care, then the practice team may believe they are making more of a difference in patient's lives. The major finding from this study is that PVPs includes actions that all practices can incorporate to improve patients' experiences with care.

Acknowledgements

The authors would like to thank Ms. Halyna Didura for her assistance with data preparation. We also thank Drs. Rebecca Lipner, Senior Vice President, and Robin Guille, Vice President, of the Evaluation, Research and Development Department at the American Board of Internal Medicine for their critical but helpful comments. All authors contributed to the study conception and design; to the acquisition, analysis, and interpretation of the data; and drafted these proceedings. There are no funders for this

paper but the authors are employees of the American Board of Internal Medicine. There were no prior presentations of this work.

References

AHRQ (2012) Patient Experience Measures from the CAHPS® Clinician & Group Surveys Document No. 1309 Updated 05/01/2012

American Society for Quality (2014) Plan-Do-Check-Act (PDCA) Cycle http://asq.org/learn-about-quality/project-planning-tools/overview/pdca-cycle.html

Arnold G, Baranowski R, and Duhigg L (2013) Can Planned Visits Improve Provider-Patient Interactions? A Case-Control Study Presentation for the 141th Annual Meeting and Exposition of the American Public Health Association, November 2013, Boston, MA

Austin PC (2010) The Performance of Different Propensity-score Methods for Estimating Differences in Proportions (Risk differences or Absolute risk reductions) in Observational Studies. Statistics in Medicine 29: 2137–48

Cochran WG, (1954) The Combination of Estimates from Different Experiments. Biometrics 10: 101-29.

Coleman K, Austin BT, Brach C, and Wagner EH (2009) Evidence On The Chronic Care Model In The New Millennium Health Affairs, 28, no.1:75-85.

DerSimonian R and Laird N (1986) Meta-analysis in Clinical Trials. Controlled Clinical Trials 7: 177-88.

Edgman-Levitan, S, Shaller D, McInnes K, Joyce R, Coltin KL, and Cleary PD (2003) The CAHPS® Improvement Guide Practical Strategies for Improving the Patient Care Experience. Department of Health Care Policy, Harvard Medical School, CMS contract No. 500-95-0057, 102-4.

Edgman-Levitan, S, Shaller D, McInnes K, Joyce R, Coltin KL, and Cleary PD (2012) The CAHPS® Improvement Guide Practical Strategies for Improving the Patient Care Experience. AHRQ, https://cahps.ahrq.gov/quality-improvement/improvement-guide/improvement-guide.html 115-19.

Glaseroff (2007) Redesigning Chronic Illness Care: Evidence, Experiences, and Stakeholders for Presentation for the IHI National Forum 2007 Orlando, Florida.

Kosanke J and Bergstralh E (2004) GMATCH SAS® Macro: Match 1 or more controls to cases using the GREEDY algorithm http://www.mayo.edu/research/documents/gmatchsas/doc-10027248

Murray DM (1998) Design and Analysis of Group Randomized Trials. Oxford University Press, New York, 140-46.

National Committee for Quality Assurance (NCQA) (2014) Physician Practice Connections® http://www.ncqa.org/Portals/0/Programs/Recognition/PPC_web.pdf

Rosenbaum, P. R & Rubin, D. B. (2006) The Central role of the propensity score in observational studies for causal effects. In Matched Sampling for Causal Effects 2006 Cambridge University Press NY reprinted from Biometrika, 1983, 70 Pages 41-55.

Rosenthal R, Rosnow RL, and Rubin DB (2000) Contrasts and Effect Sizes in Behavioral Research: A Correlational Approach. Cambridge University Press, Cambridge, UK 8-36.

Sheskin DJ (2004) Handbook of Parametric and Nonparametric Statistical Procedures (3rd ed.) Chapman/Hall/CRC Baton Rouge FL. 1061-78

Sutton AJ, Abrams KR, Jones DR, Sheldon, TA, and Song F. (2000) Methods for Meta-Analysis in Medical Research. Wiley & Sons, Chichester, West Sussex, England, 93-105.

Wagner (2007) Redesigning Chronic Illness Care: The Chronic Care Model Presentation for the IHI National Forum 2007 Orlando, Florida.
Zar JH (1999) Biostatistical Analysis (4th ed.) Prentice Hall, Upper Saddle River, NJ.