APPENDIX A. SEARCH STRATEGIES

Main Search: March 14, 2017

(((((((((("Minority Health"[Mesh]) OR ("Healthcare Disparities"[Mesh] OR "Health Status Disparities"[Mesh])) OR "Racism"[Mesh]) OR "Sexism"[Mesh]) OR "Homophobia"[Mesh]) OR "Observer Variation"[Mesh])) OR ((((((unintended[Title/Abstract] OR unanticipated[Title/Abstract] OR unforeseen[Title/Abstract] OR Unexpected[Title/Abstract] OR unplanned[Title/Abstract] OR undesired[Title/Abstract] OR unconscious[Title/Abstract])) OR exception reporting[Title/Abstract]) OR preferential reporting[Title/Abstract]) OR bias[Title/Abstract]) OR (gaming[Title/Abstract] OR spillover[Title/Abstract] OR "risk selection"[Title/Abstract] OR unincentivised[Title/Abstract] OR nonincentivised[Title/Abstract])))) AND ((("Physician Incentive Plans"[Mesh]) OR "Reimbursement, Incentive"[Mesh]) OR "economics"[Subheading]) AND VA filter: (("Veterans Health"[Mesh])) OR (((VA OR Veteran OR VAMC OR Veterans)) OR ("Veterans"[Mesh] OR "United States Department of Veterans Affairs"[Mesh] OR "Hospitals, Veterans"[Mesh]))	
Limited to Publication Date after 1/1/2014	
PubMed	N = 107
CINAHL (after deduplication with PubMed search)	N = 34
PsycINFO (after deduplication with PubMed and CINAHL)	N = 116

Author Search: March 28, 2017

(((((Petersen L[Author]) OR Kerr E[Author]) OR Hofer T[Author]) OR Benzer J[Author]) OR Werner R[Author]) OR Volpp K[Author]	
PubMed	N = 618

APPENDIX B. TECHNICAL EXPERT PANEL AND KEY INFORMANTS

TECHNICAL EXPERT PANEL MEMBER AND KEY INFORMANT

Justin Benzer, PhD
Research Health Scientist; Department of Veterans Affairs
Associate Professor; University of Texas at Austin Dell Medical School
Research Associate Professor; Public Health, Texas A&M University

Laura A. Petersen, MD, MPH, FACP
Associate Chief of Staff for Research; Michael E. DeBakey VA Medical Center
Director; Center for Innovations In Quality, Effectiveness, and Safety (IQuESt)
Professor of Medicine, Chief of Section of Health Services Research; Baylor College of Medicine

Laura Damschroder, MS, MPH
Investigator; Center for Clinical Management Research, VA Ann Arbor Healthcare System

Rachel Werner, MD, PhD
Core Investigator; VA Center for Health Equity Research and Promotion (CHERP)
Professor of Medicine; Division of General Internal Medicine, University of Pennsylvania Perelman School of Medicine
Director of Health Policy and Outcomes Research, Medicine; University of Pennsylvania Perelman School of Medicine
Senior Fellow; Leonard Davis Institute of Health Economics, University of Pennsylvania
Attending Physician; Philadelphia VA Medical Center

KEY INFORMANTS

Alvaro Sanchez, MD
Chief Medical Officer; VA Midwest Healthcare Network (VISN 23)

Gary Young, JD, PhD
Director; Northeastern University Center for Health Policy and Healthcare Research
Professor of Strategic Management and Healthcare Systems; Northeastern University

Hallie Prescott, MD, MSc
Assistant Professor; Department of Internal Medicine, University of Michigan Medical School

Ilse Wiechers, MD, MPP, MHS
Assistant Professor of Psychiatry; Yale University School of Medicine
Associate Director of the Northeast Program Evaluation Center; VA Office of Mental Health Operations (OMHO)
National Program Director; Psychotropic Drug Safety Initiative

Jeff Kullgren, MD, MS, MPH
Research Scientist; VA Center for Clinical Management Research, VA Ann Arbor Healthcare System
Assistant Professor; Department of Internal Medicine, University of Michigan Medical School and Institute for Healthcare Policy and Innovation

Jeremy Sussman, MD, MS
Research Scientist; Center for Clinical Management Research, VA Ann Arbor Health System
Assistant Professor; Department of Internal Medicine, University of Michigan Medical School

James F. Burgess, Jr., PhD
Senior Investigator; HSR&D's Center for Healthcare Organization and Implementation Research (CHOIR)
Director of the Health Economics Program; Boston University

Joan E. McInerney, MD, MBA, MA, FACEP
Network Director; VISN 2

Praveen Mehta, MD
Chief Medical Officer; VISN 12

Scott Pawlikowski, MD
Primary Care Lead Physician; VA Great Lakes Healthcare System, VISN 12

Steven D. Pizer, PhD
Director of Health Care Financing and Economics; VA Boston Healthcare System
Associate Professor; Department of Pharmacy Practice and Department of Economics, Northeastern University

Sylvia J. Hysong, PhD
Investigator and Ph.D. Fellowship Program Director; Health Policy, Quality & Informatics Program, Center for Innovations in Quality, Effectiveness and Safety, Michael E. DeBakey VA Medical Center
Associate Professor; Department of Medicine, Section of Health Services Research, Baylor College of Medicine
Adjunct Assistant Professor; University of Houston

Tim Hofer, MD, MSc
Professor; Division of General Medicine, University of Michigan
Associate Director for Analytic and Information Resources; VA Center for Practice Management and Outcomes Research

APPENDIX C. STUDY SELECTION CRITERIA – PICOTS

Key Questions	KQ1. What are the effects of pay-for-performance programs on the quality of care in veteran populations?	KQ2. In Veteran populations, what are the potential unintended consequences of pay-for-performance in healthcare?	KQ3. What metrics have been commonly incentivized in the published literature examining P4P?	KQ4. In Veteran populations, what program features and implementation factors modify the effectiveness of pay-for-performance programs?	KQ5. What novel approaches and/or current or recently closed research projects funded by the VA examine the effectiveness, implementation factors, or unintended consequences associated with pay-for-performance in Veteran populations?
Population	Healthcare providers at the individual, managerial (*eg,* VISN directors), group, and institutional levels. Veteran populations whose providers/health care systems are being incentivized based on performance metrics.				
Interventions	Pay-for-performance programs targeting providers, managers, health care systems				
Comparators	Other financial incentive models; other payment models (*eg,* fee for service, bundled payments)				
Outcomes	Utilization of healthcare services Quality of healthcare services Intermediate outcomes/process of care metrics Patient evaluations of care Patient health outcomes				
Timing	Long- or short-term				
Study design	RCTs, NRCTs. Eligible observational studies: have a comparator, ≥ 3 time points and report a trend (*eg,* ITS), or n ≥ 10,000 cross-sectional or uncontrolled before–after study	All quantitative and qualitative study designs.			
Setting	VHA, CBOCs, or community sites serving Veteran populations		VHA or healthcare systems similar to the VHA or the US	VHA, CBOCs, or community sites serving Veteran populations	

APPENDIX D. STUDY SELECTION CRITERIA – INCLUSION/EXCLUSION CRITERIA

1. Population: Is the study population Veterans?
   Yes.....Proceed to #2
   No ..... Code **X1 for KQ1**. STOP

2. Financial Incentives: Does the article report outcomes and report outcomes separately for groups with and without financial incentives at the provider, managerial, group, or institutional level.
   Yes ..... Proceed to #3
   No .....Code **X2 for KQ1.** Add code **B** if retaining for background/discussion. STOP

3. KQ2: Does the article discuss or report implementation factors that modify the effectiveness of financial incentives?
   Yes .....Code **I for KQ2**. Proceed to #4
   No .....Code **X for KQ2.** Proceed to #4

4. KQ3: Does the article report unintended consequences/health disparities related to financial incentives?
   Yes .....Code **I for KQ3**. Proceed to #5
   No .....Code **X for KQ3.** Proceed to #5

5. Outcomes Does the article report utilization, quality of care (*eg,* intermediate, patient evaluations of care), or patient health outcomes (modeling studies are not included)?
   Yes ..... Proceed to #6
   No .....Code **X3 for KQ1.** Add code **B** if retaining for background/discussion. STOP

6. Publication Type: Does the article present original study data, a systematic review, or meta-analysis? Narrative or non-systematic reviews, letters, editors, and commentaries are excluded.
   Yes ..... Proceed to #7
   No .....Code **X4 for KQ1.** Add code **B** if retaining for background/discussion. STOP

7. Systematic Review: Is the article a systematic review or meta-analysis of primary studies?
   Yes .....Code **SR for KQ1**. STOP
   No .....Proceed to # 8

8. Case Studies/Case Series: Does the article present a case study, case series, or case report?
   Yes .....Code **X5 for KQ1**. STOP
   No .....Proceed to # 9

9. Comparator/Study design: Is the article a primary study that compares a financial incentive to another financial incentive model or no financial incentive/usual care, or does it report 3 or more time points (and trend data), or have more than 10,000 participants
   Yes ..... STOP
   No .....Code **X7 for KQ1**. Add code **B** if retaining for background/discussion. STOP.

APPENDIX E. KEY INFORMANT INTERVIEW GUIDE

Portland Evidence-based Synthesis Program

Pay-for-Performance and Veteran Care: Effects, Implementation, and Unintended Consequences

Introduction

- § ESP
- § Project
- § Project Description

1. Given your experience, what factors do you think are most important for the VA to consider in implementing pay for performance programs?

   b. Can you give an example of a situation where [the factor/s mentioned above] made a difference in the success/failure of implementing a P4P program within VA?

   c. Do you think factors influencing implementation success of P4P in the CHOICE Program would be similar to, or different, from those you mentioned above? Why?

Next I want to talk more specifically about aspects of P4P programs that you believe are important.

Measures

2. Are there types of measures that you believe the VA should prioritize?
   a. Can you give us an example of such a measure?

3. Are there types of measures you have found problematic/think might be problematic within a VA context?
   a. Can you provide an example?

4. How might measures differ when used in the context of P4P versus the CHOICE Program?

Incentives

5. Can you tell us about different types of incentives that you think are important (*eg*, rewards vs penalties, type/nature, frequency/duration, certainty)?

6. Do you think incentives used in the context of the CHOICE Program should differ from those used within VA? Why or why not?

Implementation factors

7. What other implementation factors should the VA consider as they set up partnerships with the community?

Probe: Implementation processes:

- measure monitoring/evaluation,
- incentive removal,
- stakeholder engagement

8. How should the VA engage stakeholders as they start setting up partnerships in the community?

Probe:

- Inner setting (Institutional)
- Outer setting (economic, political, social contexts)
- Providers
- Cognitive/affective responses
  - beliefs, attitudes
  - cognitive response constructs such as biases, professionalism, heuristics, identification with one's organization
  - behavioral response constructs such as risk selection, gaming, systems improvement responses

Unintended Consequences

9. As the VA moves forward with implementing P4P in the community, how can they minimize negative unintended consequences?

   - Examples:
     - Risk selection
     - Deterioration of un-incentivized care
     - Impairment of intrinsic motivation/professionalism
     - Gaming
     - Teaching to the test and/or overtreatment

10. Theoretically, community care can be very costly – and with a fixed budget, do you think it could potentially affect health disparities in Veteran populations? If yes, how so? Are these 2 populations likely to be different? And is there anyone working on this?
    - Examples:
      - Low income
      - Racial/ethnic minorities
      - Rural/distance from VHA
      - Homeless
      - Mental health
      - Disabilities
      - Women
      - LGBT

11. Do you know of anyone in the VA who is piloting novel performance metrics, novel approaches to P4P, etc. If so, who? (And can you tell us about these metrics or approaches?)

12. Is there anyone else you think it would be important for us to talk to?

13. Attached to the original email was a list of studies we identified for inclusion. Are there any others you suggest?

14. Is there anything else that you think it is important that we know?

APPENDIX F. RISK OF BIAS ASSESSMENT

Risk of Bias of Randomized Controlled Trials

Trial	Random sequence generation	Allocation concealment	Blinding of participants and personnel	Blinding of outcome assessment	Incomplete outcome data	Selective reporting
Petersen et al, 2013,[17] 2016[16]	Low	Low	Low	Low	Unclear	Unclear

Risk of Bias of Cohort Studies

Study	Representative-ness of the exposed cohort	Selection of the non-exposed cohort	Ascertain-ment of exposure	Description of concurrent QI Initiatives	Demonstration that the outcome of interest was not present at the start of the study	Comparability of cohorts on the basis of the design or analysis	Assess-ment of outcome	Follow-up long enough for outcomes to occur	Adequacy of follow up cohorts
Beard et al, 2013[21]	Yes	NA	Yes	No	Yes	Yes	Yes	Yes	Yes
Benzer et al, 2014[14]	Yes	NA	Yes	Yes	NA	NA	Yes	Yes	Yes
Harris et al, 2015[15]	Yes	NA	Yes	Yes	Yes	NA	Yes	Yes	NA
Hysong et al, 2011[68]	Yes	NA	Yes	No	NA	NA	Yes	Yes	Yes
Kerr et al, 2012[22]	Yes	NA	Yes	No	Yes	Yes	Yes	Yes	Yes
Petersen et al, 2009[23]	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	NA
Prentice et al, 2014[72]	Yes	Yes	Yes	No	Yes	Yes	No	Yes	NA
Rosen et al, 2016[70]	Yes	NA	Yes	Yes	Yes	Yes	Yes	Yes	NA
Saini et al, 2014[25]	Yes	NA	Yes	No	Yes	Yes	Yes	Yes	Yes

Risk of Bias of Cross-sectional Studies

Study	Representative-ness of the sample	Sample size	Non-respondents	Ascertain-ment of exposure/risk factor	The subjects in different outcome groups are comparable, based on the study design or analysis. Confounding factors are controlled.	Assessment of the outcome	Statistical test
Finley et al, 2017[76]	Yes	No	No	Yes	Yes	Yes	Yes
Frakt et al, 2017[66]	Yes	Yes	No	Yes	Yes	Yes	Yes

Hysong et al, 2012[73]	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Hysong et al, 2016[67]	Yes	Yes	NA	Yes	NA	Yes	Yes
Petersen et al, 2005[69]	Yes	Yes	NA	Yes	Yes	Yes	Yes
Prentice et al, 2016[71]	Yes	Yes	NA	Yes	Yes	Yes	Yes
Saini et al, 2016[24]	Yes	Yes	NA	Yes	Yes	Yes	Yes
Urech et al, 2015[18]	Yes	Yes	NA	Yes	Yes	Yes	Yes

APPENDIX G. PEER REVIEW DISPOSITION TABLE

Reviewer Number	Comment	Response
Are the objectives, scope, and methods for this review clearly described?		
1, 2, 3, 7	Yes	Thank you.
Are there any published or unpublished studies that we may have overlooked?		
1	No	Noted
2	Yes - You cite some presentations that led at least partly to publications: Frakt, Austin B., Jodie Trafton, and Steven D. Pizer. "The association of mental health program characteristics and patient satisfaction." The American journal of managed care 23.5 (2017): e129. Prentice, Julia C., Michael L. Davies, and Steven D. Pizer. "Which outpatient wait-time measures are related to patient satisfaction?." American Journal of Medical Quality 29.3 (2014): 227-235.	Thank you. We have replaced the presentation Frakt et al, 2016 with the suggested Frakt et al, 2017 publication, and we have added Prentice et al, 2014.
3	No	Noted
7	Yes - Hysong SJ, SoRelle R, Broussard Smitham KK, and Petersen LA (in press). Reports of Unintended Consequences of Financial Incentives To Improve Management of Hypertension. PLoS One. Hysong SJ, Knox MK, Haidet P. Examining clinical performance feedback in patient-aligned care teams. Journal of general internal medicine. 2014 Jul 1;29(2):667-74. Hysong SJ, Broussard Smitham KK, SoRelle R, Knox MK, Amspoker AB, Hughes A, Haidet P. Mental Models of Audit and Feedback in Primary Care Settings. In preparation, (target journal, Implementation Science)	Thank you. The in press article was published prior to the finalization of our report; thus, it was formally added to our synthesis. We also added Hysong et al, 2014. As the manuscript examining mental models remains unpublished, we added a brief description of the purpose to our section that describes unpublished work.
Is there any indication of bias in our synthesis of the evidence?		
1, 2, 3, 7	No	Noted
Additional suggestions or comments can be provided below. If applicable, please indicate the page and line numbers from the draft report.		
1	This is a clearly written sequel to 2 prior reviews of P4P conducted by this ESP Center. It frames the question in terms of how the findings might apply to the Veterans Choice Program. While it hardly presents earth-shattering conclusions, it represents the literature fairly and provides a timely refresh of published studies that one hopes will inform policymaking, if not immediately, hopefully sometime in the future. The questions that are unanswered are those that remain unstudied - how do health outcomes change? What is the impact on physician morale esp burnout? beyond anecdote and qualitative assessments, how prevalent and important are the unintended consequences? Hopefully some of these questions will be addressed in future HSR&D funded research.	Noted, thank you.

Reviewer Number	Comment	Response
2	The content of the report is well organized and presented, but the text needs thorough proofreading and editing for grammar and clarity with issues on almost every page. These typically do not detract seriously from the content, but combine to undermine the credibility of the report.	Noted. We have thoroughly proofread and edited the report.
2	Regarding substance, there is some tension between the idea of targeting incentives to providers instead of to systems or clinics and using population-based metrics. We want to target providers to improve the chance they will respond to the incentive, but we want population-based metrics to improve precision and encourage a population-based approach to care. It seems to me that this tension should at least be acknowledged.	Thank you. We have added a statement clarifying that population-based methods are unlikely to motivate behavior at the provider level.
2	The boxed quote on page 56 is potentially very confusing. Perhaps the key informant intended to say "I'm ordering a very expensive cat scan," but s/he said "valuable," which suggests undertreatment instead of overtreatment.	Thank you for pointing that out. We have deleted “valuable” and replaced it with “(expensive).”
3	This is an excellent document that I think will be quite helpful to VA as they try to apply their strengths to community care, where may of those strengths will be less usable. While structured in 5 key questions, there were really 2 projects here. In the first, they essentially described the research on P4P in VHA historically. In the second, they interviewed VA researchers who could help inform how VHA should approach this going forward.	Noted, thank you.
3	Both projects were well-performed and their results were clearly presented. For a long document and a slightly diffuse project, I thought the document itself was readable and the findings were clear. While there are a few research decisions that I might have made differently, it was clear what decisions they had made and why.	Noted, thank you.
3	For the question about how P4P has worked in VHA in the past, the answer was that the research isn’t very clear, they’ve used a lot of different measures, and there have been some unintended consequences, particularly overuse. The research has shown particular reliability questions with a clear rise in metric gaming and “denominator management.”	Noted.
3	For the question of how P4P can work in VCA and in the future, I thought the interviewees overall gave a fairly large number of potentially useful ideas. The most important of these is to essentially remember that P4P choices are intended as tools to help implement whatever is important to the organization. The major initial concerns about VCA are probably more about access, coordination with VA providers, and making sure community providers access the services VA truly excels at (like PTSD and rehab care), than more traditional measures like blood pressure control. Relatedly, P4P should always be seen as one tool in an integrated implementation system that would also include decision support, education, and audit and feedback, among others.	Noted.
3	My biggest problems with the document are not really addressable at this stage. My major problem was that the review work really focused on P4P in VA. However, if the question is how to use P4P to improve VCA, this is not ideal. I’d be more curious how individual insurers with non-dominant market shares have attempted to use P4P in the wild west of the community than how the centralized, mission-driven VA providers have used it. The difficulties of P4P in VCA will resemble those of private	Thank you. Given time limitations (this project had a six-month timeline) and the system-level and cultural differences between the VHA and the private sector, we focused on the VHA as a system due. We agree that

Reviewer Number	Comment	Response
	insurers more than that of internal VA historical issues.	future research should examine the issue of insurers with small market shares.
7	p. 45, lines 52-53. The goal commitment study referenced did find no difference in goal commitment between incentivized and non-incentive physicians. More importantly, though, they found goal commitment to be modest at best, suggesting there were stronger situational factors that were impacting both physician types.	Thank you. We have edited the statement to read, “One study examined a provider affective/cognitive response, and found that not only did P4P have no impact on goal commitment, but that physicians may perceive an external locus of control for hypertension care.” To the detailed findings in Table 9, we have added “In addition, patient non-adherence and inconsistent follow-up were cited as barriers to care.”
7	Figure 3. The Hysong et al. (in press) study referenced in the previous question provides new findings that many of the commonly discussed unintended consequences reflect concerns, rather than actual instances of unintended consequences occurring.	Thank you. Figure 3 highlights key informant themes, so we did not add findings from the recently published study to the figure. However, we have added the study to the report and discuss its findings.
7	P. 21, Lines 29-31. Another way to potentially mitigate overtreatment is to incentivize appropriate care rather than treatment. In other words, follow a patient over time and see whether the series of decisions made for that patient led to patient improvement (including choosing not to treat). It would be more labor intensive but would not only mitigate over treatment, it would likely have more face validity with the clinicians.	Thank you. We have added the following statement to the summary and discussion section on overtreatment: However, it is possible that overtreatment may be mitigated by incentivizing appropriate care, rather than treatment or targets, as demonstrated by the single included RCT.
7	Table 2 (p. 18) line 43 -- Peterson should be spelled "Petersen"	Thank you. Corrected.
7	P. 15 -- Rating the body of evidence; did the reviewers only include high evidence strength studies? If not, what was the criteria for including the others? Too late now, but perhaps a better approach would be to code specific characteristics of the evidence in order to better assess the strength of the evidence. This method is generally considered stronger than blanket ratings of strength.	Thank you. In addition to rating the strength of evidence for the question of effectiveness (KQ1), we did quality assess all studies using the Cochrane Risk of Bias tool for RCTs and the Newcastle Ottawa Scale for observational studies. Detailed ratings can be found in Appendix F. Given that much of the report was descriptive, other than in our methods section, we did not refer to the ratings in the body of the report.