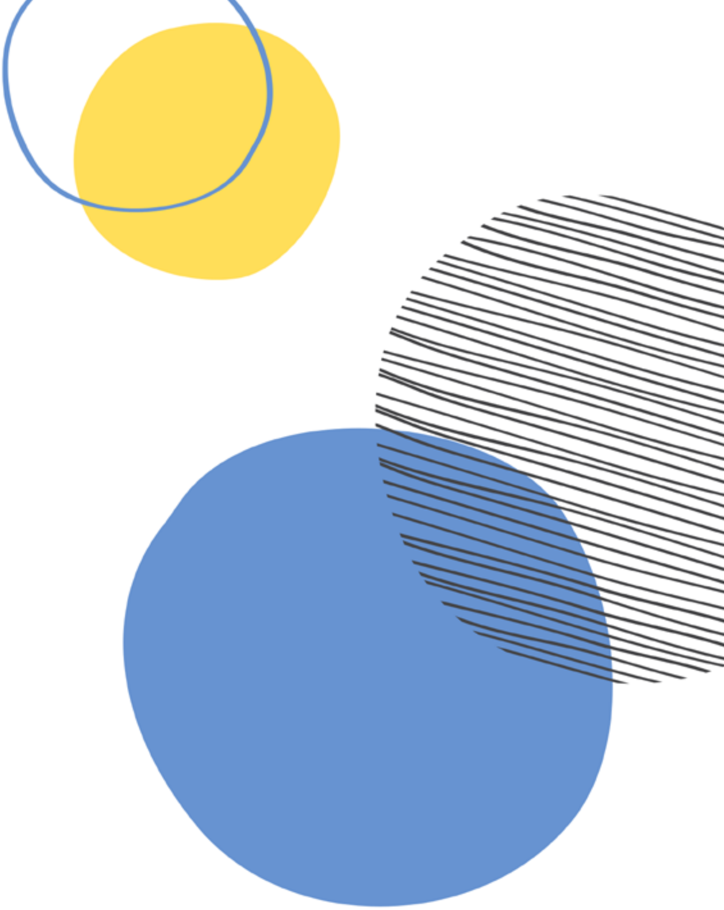


REINIE G. GERRITS

OPEN TO INTERPRETATION?

RESPONSIBLE INTERPRETATION AND REPORTING OF HEALTHCARE STATISTICS AND HEALTH SERVICES RESEARCH

OPEN TO INTERPRETATION?

Responsible interpretation and reporting of healthcare statistics and Health Services Research

COLOFON

Open to interpretation? Responsible interpretation and reporting of healthcare statistics and Health Services Research, PhD thesis.

ISBN: 978-94-028-2063-8

Cover design: Alyssa Goh

Layout: Lize Jansen

Photo of the author: Christel Egberts

Printing: Ipskamp Printing

The studies in this thesis were conducted at the Department of Public Health at Amsterdam UMC, location AMC, University of Amsterdam, Amsterdam Public Health Research institute, Amsterdam, the Netherlands. Part I of this thesis was part of the project “Statistics on Health and Healthcare” funded by the Dutch National Institute for Public Health and the Environment (RIVM). Part II of this thesis was part of the research program “Fostering Responsible Research Practices” funded by the Netherlands Organisation for Health Research and Development funded (ZonMw) grant number 445001003.

Financial support by the Academic Medical Center for the publication of this thesis is gratefully acknowledged.

OPEN TO INTERPRETATION?

Responsible interpretation and reporting of healthcare statistics and Health Services Research

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus

prof. dr. ir. K.I.J. Maex

ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen
op vrijdag 3 juli 2020, te 13.00 uur

door

Reinie Gertrudis Gerrits

geboren te Venray.

PROMOTIECOMMISSIE:

Promotor	Prof. dr. N.S. Klazinga	AMC – UvA
Co-promotores	Dr. D.S. Kringos	AMC – UvA
	Dr. M.J. van den Berg	AMC – UvA
Overige leden	Prof. dr. A.E. Kunst	AMC – UvA
	Prof. dr. L. M. Bouter	Vrije Universiteit Amsterdam
	Prof. dr. D. Ruwaard	Universiteit Maastricht
	Prof. dr. A. Verhoeff	Universiteit van Amsterdam
	Prof. dr. A.J. Pols	AMC-UvA

Faculteit der Geneeskunde

voor familie

TABLE OF CONTENTS

PART I. RESPONSIBLE INTERPRETATION OF STATISTICS ON HEALTH AND HEALTHCARE

PART II. RESPONSIBLE REPORTING OF HEALTH SERVICES RESEARCH

APPENDIX

1

General Introduction

GENERAL INTRODUCTION

Directly or indirectly, healthcare statistics and Health Services Research have a lasting influence on society. To foster responsible interpretation and reporting of research findings to policy makers, professionals and the public, a reflection of the conduct and communication of research is required. This thesis first explores the interpretation of publicly reported statistics and proposes a method to improve the interpretation of publicly reported statistics on health and healthcare. Second, this thesis explores the responsible reporting of Health Services Research in scientific and societal publications. These topics are addressed using a combination of quantitative and qualitative data collected through semi-structured interviews, focus groups, document analyses, and surveys amongst researchers, policy makers, science journalists/communicators, knowledge integration specialists, and students.

The role of scientific evidence in society

Scientific knowledge is obtained by systematically describing, organizing, registering, understanding, and explaining phenomena. In particular, the researcher aims to predict new developments, and through this prediction, control or influence these phenomena (1). As a group, researchers form a "scientific forum" that sets the methodological rules for creating scientific knowledge and insights (2). Through this endeavour, researchers strive to identify objective knowledge that can be used to advance society. Scientific knowledge informs rational decision-making processes in society (3-5). Science is considered to underlie better, objective, and more effective decisions (6). Governments, industries, and foundations are investing significantly in scientific research to advance society (7, 8). Scientific knowledge allows decision-makers to manage, monitor, and exercise control over their respective fields (5). Yet, this rationality has recently been subjected to some pushback. Terms such as "fact free politics" (9), "science as an opinion" (10), and "alternative facts" (11) reveal an uneasiness in society regarding the credibility of the role of science in politics and policy making (12, 13).

This uneasiness is not entirely irrational. Many philosophers have reasoned that truth and knowledge are unattainable or relatively situated (14). Scientific knowledge is rarely as objective as researchers or decision-makers present it. The scientific evidence that shapes knowledge is context dependent (15) and inevitably incomplete (16). To gain a strong level of scientific confidence, scientific evidence needs to be built up over time. However, because of the time and methodological constraints of research, this level of confidence is often difficult to achieve (17). Science itself is also not value-free. Scientific institutions are not neutral and scientists are not without self-interest (6, 18). Research institutions and funders of scientific research may inadvertently influence the interpretation of findings to fit their own interests. Additionally, researchers themselves may (un)consciously report conclusions that do not accurately reflect their findings based on their own convictions and beliefs (19).

The integration of scientific evidence in policy and practice is also not straightforward. Decision-making is not a cyclical process that smoothly flows from acquisition of evidence to application and, ultimately, full implementation. Instead, the process is largely iterative and context bound, and it often involves the consideration of insufficiently supported values, ideology, practicability, complexity of the subject, timeliness, and the distribution of power in politics (20-22). Evidence shifts opinions and affects decisions of decision-makers (i.e., policy makers, professionals, and the public) through a more indirect process (23). When disseminated to society, research becomes part of a greater political and societal debate (24-26). Outside the research community, most people take note of research findings through non-scientific societal publications such as press releases, newspapers, social media, internet postings, or non-scientific professional journals (26-31). The public broadcasting of research evidence influences public opinion on policy, which, in turn, affects decisions made in policy and healthcare practice (23, 32, 33). At each point of the evidence dissemination process, research is constantly reinterpreted and rephrased, with risk for misinterpretation and misreporting of scientific evidence rising substantially as well (15, 34-39). Misinterpreted evidence can still be convincing to decision-makers and potentially misdirect the advancement of practice and policy. Directly or indirectly, research has a lasting influence on society. To foster the responsible interpretation and reporting of research findings to decision-makers, a continuous reflection of the conduct and translation of research to practice is required (40, 41).

Particularly in the field of health and healthcare, decisions on policy and practice are often grounded in scientific evidence (6, 42-44). The formation of accessible, qualitative, affordable healthcare can greatly benefit from scientific research (18, 45). Many advances in healthcare were supported by the progress of scientific knowledge (24, 25, 46), including the implementation of Medicare in the USA (47) and the development of the primary healthcare system in the Netherlands (48). Studying how users interpret scientific data and statistics on health and healthcare and how researchers report Health Services Research (HSR) in scientific and societal publications could facilitate increased support for the responsible interaction between research and society.

The field of Health Services Research

HSR often meets broad societal interest, covering topics such as co-payments, evaluation of quality improvement efforts, cost-effectiveness of medications, patient empowerment, compliance with therapy and effects of policies (17).

HSR is defined as "the multidisciplinary field of scientific investigation that studies how social factors, financing systems, organizational structures and processes, health technologies, and personal behaviours effect access to healthcare, the quality and cost of healthcare and ultimately our health and well-being." (49). The field has a particular connection to policy and practice, for which providing knowledge for direct application to healthcare has been a primary purpose of the research field (47). A more defined definition of HSR has developed over the years, but the boundaries of the field expand and contract as it interacts with other scientific disciplines, such as the fields of biomedicine, economics and sociology (47).

Internationally, the field was established in the 1970s (50), but compared to the biomedical field, HSR is relatively young. The field is still fully in development (51, 52). HSR is often context bound, in which conclusions drawn are often limited to the study setting. With a direct connection to healthcare practice and policy, HSR is funded by governmental and healthcare organisations who intend to apply the knowledge in practice. Because they intent to apply evidence in policy or practice, funders are frequently involved in the progress of the study. They are thus often part of the design or interpretation (53). Qualitative, quantitative and mixed methods studies are all used in the field, often applying observational or explorative designs. Whereas much research in other scientific fields takes place in a lab or other strictly controlled setting, HSR usually takes place in real-life settings and is based on data coming from such "natural laboratories". Research methodologies in HSR are less strictly controlled as compared to the biomedical golden standard of empirical science, the randomized controlled trial (54). Fewer strict structures and norms guiding HSR studies exist. This provides an increased freedom for researchers in conducting HSR and reporting the resulting findings and conclusions and gives researchers more freedom to place results in context.

The Application of Health Services Research in Policy and Practice

Although creating evidence for policy and practice is a central goal of the field of HSR, how HSR evidence is applied in policy and practice remains difficult to determine (55).

HSR follows the trend set by the emergence of Evidence Based Medicine. Evidence Based Medicine arose nearly 30 years ago to improve the use of evidence in clinical practice (56). For instance, the Grading of Recommendations Assessment, Development and Evaluation (GRADE) guidelines were developed to determine the quality of evidence and strength of recommendations (57). The field has since expanded its boundaries to encompass all fields of healthcare, with the development of Evidence Based Practice and Policy (EBP), covering HSR as well (51, 58). The first step in EBP was to systematically review evidence and develop guidelines and measures for evidence-based implementation (51). New techniques for appropriate synthesis of policy and management relevant evidence have since been suggested (17). EBP researchers have further dedicated their efforts towards "bridging the gap" between research and practice. Throughout the past decade, barriers for evidence-based policy and practice have been identified, including a lack of collaboration between researchers and decision-makers, time constraints and inaccessible sources (59). Additionally, numerous strategies, tools, and frameworks have been designed to aid researchers and decision-makers in forming evidence-based policies (33).

To improve the connection between HSR and policy and practice across the entire field of HSR, researchers are stimulated to spread their findings via societal publications to decision-makers (60). Yet, this focus on maximizing the use of evidence in a decision-making system that needs to consider many competing values, lacks practical consideration (61). EBP has reached a point where contextualisation of evidence is neglected (40). Critics of the evidence-based health movement have noted the necessity for better contextualisation and evaluation of evidence. Moreover, the interpretation and correct translation of evidence towards society needs to be advanced (41). For instance, a smoking prevention program proved effective in one area is not necessarily effective in another with a different culture and population structure. In order to support the advancement of evidence-based policy and practice of health and healthcare, a constant inquiry is required in the interaction between scientific production and its application to society.

Although the "bridging of the gap" between research, policy and practice remains an ongoing process, we should start to look further at not only *how* but also at *what* evidence is actually transferred to policy and practice, as well as critically examine the nature of the translated scientific findings (51, 55).

Interpretation of Statistics on Health and Healthcare

Decision-makers are constantly confronted with HSR evidence in their decision-making processes and continuously challenged to interpret and apply evidence correctly (62), dealing with the context in which evidence is constructed and to which it is restricted (24).

When interpreting research, a decision-maker will synthesize information and evaluate it. In doing so, the content of information will need to be assessed. They will need to decide whether the information is relevant, appropriate, applicable, acceptable and useful for their goals and how it relates to what they already know or believe. Their evaluation of evidence is influenced by many factors, including professional or personal conflicts of interest, their ability to understand the information, their critical reflection skills, their beliefs, intuition and assumptions (32). Also, the time and effort they are willing to invest will affect their interpretation. Because of these factors, the content of information can be misinterpreted, or be applied in an unsuitable context.

Based on statistics on health and healthcare, the severity, magnitude or impact of a health issue in society is determined (63, 64), but these statistics are often subject to misinterpretation. Statistics provide the authority of the scientific community and give the user power to persuade others (8). Throughout history, governments have invested significantly in the development of data infrastructures (65). Because statistics are seemingly easy to interpret, they enable decision-makers to evaluate healthcare decisions and gain insight into the provision of accessible, high quality and affordable healthcare services (66, 67). But regardless of the deceptive clarity of statistics, rather than an objective reflection of the reality, statistics are the result of a series of decisions on subjects, inclusion of populations and restrictions in research methodology. Those who use statistics for decision making often lack insight in how these statistics were constructed and should be interpreted (59). Moreover, multiple available sources of information may report contradicting statistics on the same topic. Contradictory or unclear reporting may cause uncertainty regarding a health(care) topic (68), create a barrier (69, 70), lead to avoidance of decision making, or even contribute to misinformed decisions (67, 71). In addition, statistics on similar topics leave room for political and opportunistic use, such as applying the figure that fits best with the agenda of its users (67, 72).

A structured assessment tool may support users of statistics in their interpretation. For a long time, lay checklists have been published in the form of popular literature, such as Darrel Huffs book “How to Lie with Statistics” (73). And although many checklists for the assessment of statistics in scientific publications are available, there is a lack of evidence-based tools that can aid the practical interpretation of publicly reported statistics. Decision-makers or journalists may wonder if they can trust certain statistics, but may lack a full grasp of all aspects that need to be considered when appraising that statistic. A need exists for an easy to use tool that supports users to gain insight into the key characteristics that contribute to better interpretation of publicly reported statistics on health and healthcare (74, 75).

Reporting of Health Services Research

To adequately interpret HSR findings, messages and conclusions should be reported in a clear and responsible manner. Researchers have much freedom in the interpretation of their research findings (76). Although contextualizing findings is inherently part of the scientific method, the risk does arise that conclusions are not adequately supported by the research findings. Estimates on exaggerated reporting in biomedical research vary from 10% of publications describing conclusions discordant from study results to 100% of publications containing rhetorical practices resulting in spin (defined as *"specific intentional or unintentional reporting that fails to faithfully reflect the nature and range of findings and that could affect the impression the results produce in readers"*) (19, 77). Because misreporting of research findings occurs largely in biomedical literature (19) and HSR follows a comparable publication process, it is highly likely that HSR literature suffers from similar problems.

When scientific misreported findings are further communicated to the society and the policy world by communication officers, journalists, or even researchers themselves, they might end up in public debate. And because of the often-necessary simplification of scientific findings for policy and practice (15), inconsistencies between the original scientific publication and successive the societal publication could occur (34, 78). Researchers could also fail to accurately communicate their scientific findings towards a broader public. From the biomedical field, we know that unjustified causal claims are introduced in 20% to 33% of press releases, and that 40% of news articles give more explicit health advice to the readers than was expressed in the underlying scientific publication (34-38). The practice-oriented goal of HSR amplifies the importance of accuracy in all messages and conclusions relayed in societal publications (79). But we have little knowledge on how HSR is reported in societal publications. Therefore, more insight in the translation of findings from scientific to societal publications is needed.

To support responsible scientific reporting, the process of scientific publication is highly structured and supported by many publication checklists for a wide range of research methods (80). Scientific journals require nuance and clarity when reporting content. Most importantly, the peer-review system is built to assure that scientific publications contain necessary information regarding methodology and context of evidence. This extensive system is all but bullet-proof (81). Researchers may be influenced by various factors to inadequately report their findings in their scientific publications. Any conclusions researchers draw are relative to their presuppositions, interests and social milieu (14). They are under pressure to publish in high-impact journals, to increase their citation scores and to attract media attention to raise their prestige and chances for future research funding and job security (7, 82-85). Pressure to influence healthcare practice might push HSR scientists to spin their results to increase the policy relevance of their findings and appeal of their concluding messages (86). Influence of funders, time constraints, publication pressure, journal policy and institutional policy are well known factors that contribute to questionable research practices in biomedical research (86-90). Nevertheless, we have little knowledge pertaining to responsible reporting, particularly in the field HSR. Furthermore, we lack information on factors that may contribute to questionable reporting of messages and conclusions.

The Netherlands

The Netherlands has had a long tradition of EBP. Over the past decades many public institutions were created to support the development and integration of research to policy and practice. The Dutch institute for public health and the environment (RIVM) plays a pivotal role in the public dissemination of policy relevant statistics on health and healthcare. The RIVM controls multiple websites in which many available statistics on the health of the Dutch population can be found. In an attempt to create clarity in the abundance of available statistics, the Dutch ministry has commissioned RIVM to develop and maintain one website that functions as the main source of information in government-related information (https://www.staatvenz.nl). Regardless of a highly advanced tradition of EBP, we have little knowledge on the actual use of statistics in the Dutch policy making context in health and healthcare (15, 41, 74, 75, 91, 92). Better understanding is needed of how evidence is used in the national policy making process and what the key characteristics are that contribute to the interpretation of a statistic on health or healthcare. To responsibly publish these statistics and allow for their correct interpretation by its users, the RIVM funded the work within this thesis on the interpretation of statistics on health and healthcare.

The Netherlands has a relatively small, homogenous HSR community. Because the community is well connected, it is expected to be possible to find consensus on measures to improve responsible reporting. The Dutch field of HSR is led by a number of academic and non-academic institutions.

Each of these institutions have their own areas of specialisation, but have strong collaborative ties to each other. All institutions conduct scientific inquiry, with some focusing on direct policy and practice interaction, primarily through scientific reports for policy and practice, although others prioritize publication in international peer-reviewed journals. Although the Dutch field holds strong connections, each institution has their own structures in place to support their researchers. This variety offers the opportunity to learn from different scientific and societal reporting experiences at these institutions.

The Netherlands Organisation for Health Research and Development (ZonMw) is the national Dutch funder of research on health and healthcare. To address the need for greater quality, integrity and efficiency in academic research, they funded the program "Fostering responsible research practices" to finance "research on research". The second part of this thesis on the reporting of HSR was funded through this program.

Aims of the thesis

We lack knowledge in the understanding of how evidence is used in the national policy making process and what key characteristics contribute to the interpretation of statistics on health or healthcare. A need exists for an easy to use tool that supports users in the interpretation of publicly reported statistics on health and healthcare. Furthermore, we have little knowledge pertaining to responsible reporting in scientific and societal HSR publications and contributing factors.

The aims of this thesis are therefore twofold:

- One, to propose a method to improve the interpretation of publicly reported statistics on health and healthcare.

- Two, to provide insight in the scientific and public reporting of Health Services Research.

Thesis outline

To address these aims, this thesis is structured in two parts: Part I, responsible interpretation of statistics on health and healthcare. Part II, responsible reporting of Health Services Research.

Part I. Responsible interpretation of statistics on health and healthcare

Part I consists of three chapters aimed at gaining insight in the interpretation of statistics on health and healthcare, and the development of a method to improve the interpretation of publicly reported statistics on health and healthcare.

Chapter 2 explores how different statistics are used in the Dutch government and parliament debates on health and healthcare through a conventional qualitative content analysis. Chapter 3 maps the key characteristics relevant to the interpretation of statistics on health and healthcare and describes the development of a Figure Interpretation Assessment Tool-Health (FIAT-Health). This tool enables a systematic assessment of statistics on health and healthcare allowing for a better interpretation of these statistics. Chapter 4 expands on this study, by testing and evaluating the FIAT-Health 1.0 amongst its intended user groups, and further refining the tool based on those results. These studies resulted in the FIAT-Health 2.0, an online qualitative appraisal tool that has the potential to aid the interpretation and public reporting of statistics on health and healthcare.

Part II. Responsible reporting of Health Services Research

Part II consists of three chapters on the responsible reporting of HSR in scientific and societal publications, and the factors that may contribute to questionable research practices (QRPs) in scientific HSR publications.

Chapter 5 examines the occurrence and nature of QRPs in the reporting of messages and conclusions in international scientific HSR publications. This chapter describes the construction and validation of an assessment instrument of possible QRPs in the reporting of messages and conclusions, and the assessment of these QRPs in scientific HSR publications authored by researchers from HSR institutions in the Netherlands. Chapter 6 studies possible inconsistencies and replicated QRPs in non-scientific societal publications based on international scientific HSR publications. Chapter 7 explores the individual, institutional and scientific environment factors potentially associated with QRPs in scientific HSR publications, as assessed in chapter 5. The construction of a comprehensive framework of factors related to QRPs is included in this chapter.

Data and data sources

Table 1 provides an overview of the studies described in chapter 2 to 7 and their respective data sources and methodologies. This thesis includes both qualitative and quantitative methodologies, reported as qualitative and mixed-methods studies. All studies were conducted with primarily Dutch participants and within a Dutch setting. The findings of this thesis are discussed and reflected upon in chapter 8.

Table 1 | Overview of chapters and respective study characteristics

Chapter	Research questions	Data source	Methodology
Part I			
2	How are different types of statistics used in various ongoing policy debates on health and healthcare in the Netherlands?	Policy documents on 8 topics derived from the Dutch parliament and government websites	Qualitative, conventional content analysis
3	(1) What are the key characteristics relevant to the interpretation of statistics on health and healthcare? (2) What method can improve the proper use of these statistics by policymakers, managers, scientists, patients, and the general public?	Expert consultation, focus groups, semi-structured interviews	Qualitative, thematic content analyses
4	(1) What is the usability of the FIAT-Health 1.0 amongst its intended user groups? (2) How can the FIAT-Health 1.0 be improved?	Test and evaluation forms of FIAT-Health 1.0, expert consultation, test session FIAT-Health 2.0	Qualitative, content analyses
Part II			
5	(1) What is the occurrence and nature of Questionable Research Practices (QRPs) in the reporting of messages and conclusions in international scientific Health Service Research (HSR) publications authored by researchers from HSR institutions in the Netherlands? (2) What is the relationship between study type, methodology and design and the occurrence of QRPs?	International scientific HSR publications authored by researchers from HSR institutions in the Netherlands	Mixed methods, assessment of scientific publications, quantitative analyses of relationship study type. methodology, and methodology
6	(1) Are societal publications on HSR consistent with the messages reported in the underlying research papers? (2) Are QRPs in scientific HSR publications replicated in societal publications? (3) Do fewer inconsistencies occur in societal publications if the first scientific author is involved in writing them?	International scientific HSR publications, and societal HSR publications that derived from them	Mixed methods, Qualitative content analysis of replicated QRPs and inconsistencies in societal publications Quantitative analyses of contribution of scientific author
7	How do individual, institutional, and scientific environment factors associate with QRPs messages and conclusions in scientific HSR publications?	International scientific HSR publications, literature review, semi-structured interviews, focus groups, and survey data	Mixed-methods, Explorative literature review, qualitative thematic analyses Quantitative analyses of survey data with assessed QRPs

REFERENCES

1. de Groot AD. Methodologie. Grondslagen van onderzoek en denken in de gedragswetenschappen. Assen: Van Gorcum; 1994. p. 19.
2. de Groot AD. Methodologie. Grondslagen van onderzoek en denken in de gedragswetenschappen. Assen: Van Gorcum; 1994. p. 27-8.
3. Gordon C. Governmental rationality: an introduction. In: Burchell G, Gordon C, Miller P, editors. The Foucault Effect Studies in Governmentality. Chicago: The University of Chicago Press; 1991. p. 2-3.
4. Head BW. Reconsidering evidence-based policy: Key issues and challenges. Policy and Society. 2010;29(2):77-94.
5. Foucault M, Lotringer S, Hochroth L. The politics of truth: Semiotext(e); 1997.
6. Weingart P. Scientific expertise and political accountability: paradoxes of science in politics. Science and Public Policy. 1999;26(3):151-61.
7. Weingart P. Science and the media. Res Policy. 1998;27(8):869-79.
8. Espeland WN, Stevens ML. A Sociology of Quantification. Arch Eur Sociol. 2008;49(3):397-432.
9. Veldheer Vr. De gevolgen van ‘fact-free politics’ voor beleid en onderzoek. The Hague / Nijmegen: Sociaal en Cultureel Planbureau; 2012.
10. Groen TP. Startdocument ‘Systeemfalen’ The Hague; 2013.
11. Speed E, Mannion R. The rise of post-truth populism in pluralist liberal democracies: challenges for health policy. Int J Health Policy Manag. 2017;6(5):249-51.
12. McKee M, Greer SL, Stuckler D. What will Donald Trump’s presidency mean for health? A scorecard. The Lancet.
13. Strassheim H, Kettunen P. When does evidence-based policy turn into policy-based evidence? Configurations, contexts and mechanisms. Evidence & Policy: A Journal of Research, Debate and Practice. 2014;10(2):259-77.
14. Papineau D. Philosophy of Science. In: Tsui-James NBEP, editor. The Blackwell Companion to Philosophy. Cambridge MA, USA: Blackwell Publishers Ltd; 1995. p. 301.
15. Greenhalgh T, Russell J. Evidence-based policymaking: a critique. Perspect Biol Med. 2009;52(2):304-18.
16. Upshur RE, VanDenKerkhof EG, Goel V. Meaning and measurement: an inclusive model of evidence in health care. Journal of evaluation in clinical practice. 2001;7(2):91-6.
17. Sheldon TA. Making evidence synthesis more useful for management and policy-making. SAGE Publications Sage UK: London, England; 2005.
18. Parkhurst J. The Politics of Evidence: From evidence-based policy to the good governance of evidence. Abingdon: Routlidge; 2017.
19. Boutron I, Ravaud P. Misrepresentation and distortion of research in biomedical literature. Proc Natl Acad Sci U S A. 2018;115(11):2613-9.
20. Davis P, Howden-Chapman P. Translating research findings into health policy. Social Science & Medicine. 1996;43(5):865-72.
21. Weiss CH. The Interface between Evaluation and Public Policy. Evaluation. 1999;5(4):468-86.
22. Dobrow MJ, Goel V, Upshur REG. Evidence-based Health Policy: Context and Utilization. Soc Sci Med. 2004;58.
23. Burstein P. The impact of public opinion on public policy: A review and an agenda. Polit Res Quart. 2003;56(1):29-40.

24. Lavis JN, Ross SE, Hurley JE, Hohenadel JM, Stoddart GL, Woodward CA, et al. Examining the role of health services research in public policymaking. Milbank Q. 2002;80(1):125-54.
25. Kuruvilla S, Mays N, Walt G. Describing the impact of health services and policy research. J Health Serv Res Policy. 2007;12 Suppl 1:S1-23-31.
26. Gerrits RG, van den Berg MJ, Klazinga NS, Kringos DS. Statistics in Dutch policy debates on health and healthcare. Health Res Policy Syst. 2019;17(1):55.
27. Turney J. Public understanding of science. Lancet. 1996;347(9008):1087-90.
28. Ritter A. How do drug policy makers access research evidence? Int J Drug Policy. 2009;20(1):70-5.
29. Zardo P, Collie A. Type, frequency and purpose of information used to inform public health policy and program decision-making. Bmc Public Health. 2015;15.
30. Dobbins M, Jack S, Thomas H, Kothari A. Public health decision-makers' informational needs and preferences for receiving research evidence. Worldviews Evid Based Nurs. 2007;4(3):156-63.
31. McVay AB, Stamatakis KA, Jacobs JA, Tabak RG, Brownson RC. The role of researchers in disseminating evidence to public health practice settings: a cross-sectional study. Health Res Policy Syst. 2016;14(1):42.
32. van de Goor I, Hamalainen RM, Syed A, Juel Lau C, Sandu P, Spitters H, et al. Determinants of evidence use in public health policy making: Results from a study across six EU countries. Health Policy. 2017;121(3):273-81.
33. Orton L, Lloyd-Williams F, Taylor-Robinson D, O'Flaherty M, Capewell S. The use of research evidence in public health decision making processes: systematic review. PLoS One. 2011;6(7):e21704.
34. Sumner P, Vivian-Griffiths S, Boivin J, Williams A, Bott L, Adams R, et al. Exaggerations and Caveats in Press Releases and Health-Related Science News. PLoS One. 2016;11(12):e0168217.
35. Schat J, Bossema FG, Numans ME, Smeets I, Burger P. Exaggerated health news. Nederlands Tijdschrift voor Geneeskunde. 2018;162(1).
36. Sumner P, Vivian-Griffiths S, Boivin J, Williams A, Venetis CA, Davies A, et al. The association between exaggeration in health related science news and academic press releases: retrospective observational study. BMJ : British Medical Journal. 2014;349.
37. Buhse S, Rahn AC, Bock M, Muhlhauser I. Causal interpretation of correlational studies - Analysis of medical news on the website of the official journal for German physicians. PLoS One. 2018;13(5):e0196833.
38. Haber N, Smith ER, Moscoe E, Andrews K, Audy R, Bell W, et al. Causal language and strength of inference in academic and media articles shared in social media (CLAIMS): A systematic review. PLoS ONE. 2018;13(5).
39. Smith R. Measuring the social impact of research. BMJ. 2001;323(7312):528.
40. Greenhalgh T, Howick J, Maskrey N, Evidence Based Medicine Renaissance G. Evidence based medicine: a movement in crisis? BMJ. 2014;348:g3725.
41. Russell J, Greenhalgh T, Byrne E, McDonnell J. Recognizing rhetoric in health care policy analysis. J Health Serv Res Policy. 2008;13(1):40-6.
42. Walshe K, Rundall TG. Evidence-based management: from theory to practice in health care. Milbank Q. 2001;79(3):429-57, IV-V.
43. Lavis J, Davies H, Gruen R, Walshe K, Farquhar C. Working Within and Beyond the Cochrane Collaboration to Make Systematic Reviews More Useful to Healthcare Managers and Policy Makers. Healthcare Policy. 2005;1.
44. Black N. Evidence based policy: proceed with care. BMJ. 2001;323(7307):275-9.

45. Grol R. Successes and failures in the implementation of evidence-based guidelines for clinical practice. Med Care. 2001;39(8 Suppl 2):II46-54.
46. Hanney SR, Gonzalez-Block MA, Buxton MJ, Kogan M. The utilization of health research in policy-making: concepts, examples and methods of assessment. Health Res Policy Syst. 2003;1.
47. Colby DC, Baker LC. Health services research in 2020: an assessment of the field's workforce needs. Health Serv Res. 2009;44(6):2193-7.
48. van Weel C, Schers H, Timmermans A. Health Care in The Netherlands. The Journal of the American Board of Family Medicine. 2012;25(Suppl 1):S12-S7.
49. Juttmann RE. Het domein van gezondheidszorgonderzoek. In: Plochg T, Juttmann RE, Klazinga NS, Mackenbach JP, editors. Handboek gezondheidszorgonderzoek. Houten: Bohn Stafleu van Loghum; 2007. p. 22-3.
50. Lohr KN, Steinwachs DM. Health Services Research: An Evolving Definition of the Field. Health Serv Res. 2002;37(1):15-7.
51. Clancy CM, Slutsky JR, Patton LT. Evidence-based health care 2004: AHRQ moves research to translation and implementation. Health Serv Res. 2004;39(5):xv-xxiii.
52. Bindman AB. The evolution of health services research. Health services research. 2013;48(2 Pt 1):349-53.
53. Pittman P. Health Services Research in 2020: Data and Methods Needs for the Future. Health Services Research. 2010;45(5 Pt 2):1431-41.
54. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. BMJ. 2004;328(7454):1490.
55. Lewis S. How has health services research made a difference? Healthcare policy = Politiques de sante. 2011;6(Spec Issue):74-9.
56. Guyatt G, Cairns J, Churchill D, Cook D, Haynes B, Hirsh J, et al. Evidence-Based Medicine: A New Approach to Teaching the Practice of Medicine. JAMA. 1992;268(17):2420-5.
57. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924-6.
58. Dobrow MJ, Goel V, Upshur REG. Evidence-based health policy: context and utilisation. Social Science & Medicine. 2004;58(1):207-17.
59. Innvaer S, Vist G, Trommald M, Oxman A. Health policy-makers' perceptions of their use of evidence: a systematic review. J Health Services Res Policy. 2002;7.
60. Council for Medical Sciences. The societal impact of applied health research: Towards a quality assessment system. Amsterdam: Royal Netherlands Academy of Arts and Sciences; 2002.
61. Cairney P, Oliver K. Evidence-based policymaking is not like evidence-based medicine, so how far should you go to bridge the divide between evidence and policy? Health Res Policy Sy. 2017;15(1):35.
62. Weiss CH. The Many Meanings of Research Utilization. Public Administration Review. 1979;39(5):426-31.
63. Williamson JML, Skinner CI, Hocken DB. Death and illness as depicted in the media. Int J Clin Pract. 2011;65(5):547-51.
64. Porter ME. What Is Value in Health Care? New England Journal of Medicine. 2010;363(26):2477-81.
65. Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, Gülmezoglu AM, et al. How to increase value and reduce waste when research priorities are set. The Lancet. 2014;383(9912):156-65.
66. Sato H. Agenda Setting for Smoking Control in Japan, 1945-1990: Influence of the Mass Media on National Health Policy Making. Journal of Health Communication. 2003;8(1):23-40.

67. Furedi A. The public health implications of the 1995 ‘pill scare’. Hum Reprod Update. 1999;5(6):621-6.
68. Schwitzer G. Addressing tensions when popular media and evidence-based care collide. Bmc Med Inform Decis. 2013;13.
69. Oliver K, Innvar S, Lorenc T, Woodman J, Thomas J. A systematic review of barriers to and facilitators of the use of evidence by policymakers. BMC Health Serv Res. 2014;14:2.
70. de Goede J, Putters K, van der Grinten T, van Oers HA. Knowledge in process? Exploring barriers between epidemiological research and local health policy development. Health Res Policy Syst. 2010;8:26.
71. Fischhoff B, Davis AL. Communicating scientific uncertainty. P Natl Acad Sci USA. 2014;111:13664-71.
72. When Evidence Collides With Anecdote, Politics, and Emotion: Breast Cancer Screening. Annals of Internal Medicine. 2010;152(8):531-2.
73. Huff D. How to Lie With Statistics. New York: W. W. Norton & Company; 1954.
74. Walker N, Bryce J, Black RE. Interpreting health statistics for policymaking: the story behind the headlines. The Lancet. 2007;369(9565):956-63.
75. Han PK, Moser RP, Klein WM, Beckjord EB, Dunlavy AC, Hesse BW. Predictors of perceived ambiguity about cancer prevention recommendations: sociodemographic factors and mass media exposures. Health Commun. 2009;24(8):764-72.
76. de Groot AD. Methodologie. Grondslagen van onderzoek en denken in de gedragswetenschappen. Assen: Van Gorcum; 1994. p. 49.
77. Ochodo EA, de Haan MC, Reitsma JB, Hooft L, Bossuyt PM, Leeflang MM. Overinterpretation and misreporting of diagnostic accuracy studies: evidence of “spin”. Radiology. 2013;267(2):581-8.
78. Waddell C, Lomas J, Lavis JN, Abelson J, Shepherd CA, Bird-Gayson T. Joining the conversation: newspaper journalists’ views on working with researchers. Healthcare policy = Politiques de sante. 2005;1(1):123-39.
79. Alla K, Hall WD, Whiteford HA, Head BW, Meurk CS. How do we define the policy impact of public health research? A systematic review. Health Res Policy Sy. 2017;15(1):84-.
80. Altman DG, Simera I, Hoey J, Moher D, Schulz K. EQUATOR: reporting guidelines for health research. Lancet. 2008;371.
81. Malički M, Aalbersberg IJ, Bouter L, ter Riet G. Journals’ instructions to authors: A cross-sectional study across scientific disciplines. PLOS ONE. 2019;14(9):e0222157.
82. Bornmann L. Measuring the societal impact of research: research is less and less assessed on scientific impact alone--we should aim to quantify the increasingly important contributions of science to society. EMBO reports. 2012;13(8):673-6.
83. Ioannidis JPA, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. The Lancet. 2014;383(9912):166-75.
84. Tijdink JK, Schipper K, Bouter LM, Maclaine Pont P, de Jonge J, Smulders YM. How do scientists perceive the current publication culture? A qualitative focus group interview study among Dutch biomedical researchers. BMJ Open. 2016;6(2):e008681.
85. Bouter LM. Commentary: Perverse Incentives or Rotten Apples? Accountability in Research. 2015;22(3):148-61.
86. Tijdink JK, de Rijcke S, Vinkers CH, Smulders YM, Wouters P. [Publication pressure and citation stress; the influence of achievement indicators on scientific practice]. Ned Tijdschr Geneeskd. 2014;158:A7147.

87. Moher D, Glasziou P, Chalmers I, Nasser M, Bossuyt PM, Korevaar DA, et al. Increasing value and reducing waste in biomedical research: who's listening? Lancet. 2016;387(10027):1573-86.
88. Fanelli D, Costas R, Lariviere V. Misconduct Policies, Academic Culture and Career Stage, Not Gender or Pressures to Publish, Affect Scientific Integrity. PLoS One. 2015;10(6):e0127556.
89. Maggio L, Dong T, Driessen E, Artino A, Jr. Factors associated with scientific misconduct and questionable research practices in health professions education. Perspectives on medical education. 2019;8(2):74-82.
90. Waaijer CJF, Teelken C, Wouters PF, Van Der Weijden ICM. Competition in Science: Links between Publication Pressure, Grant Pressure and the Academic Job Market. Higher Education Policy. 2018;31(2):225-43.
91. Murray SJ, Holmes D, Rail G. On the constitution and status of 'evidence' in the health sciences. Journal of Research in Nursing. 2008;13(4):272-80.
92. Peters E. Beyond Comprehension: The Role of Numeracy in Judgments and Decisions. Curr Dir Psychol Sci. 2012;21(1):31-5.

PART I

RESPONSIBLE INTERPRETATION OF STATISTICS ON HEALTH AND HEALTHCARE

2

Statistics in Dutch policy debates on health and healthcare

AUTHORS:
Reinie G. Gerrits, Michael J. Van Den Berg, Niek S. Klazinga,
Dionne S. Kringos

Published in Health Research Policy and Systems, 2019; 17:55

ABSTRACT

BACKGROUND The notion of 'fact-free politics' is debated in Europe and the USA and has particular relevance for the use of evidence to underpin health and healthcare policies. To better understand how evidence on health and healthcare is used in the national policy-making process in the Netherlands, we explore how different statistics are used in various policy debates on health and healthcare in the Dutch government and parliament.

METHODS We chose the following eight ongoing policy debates as case studies representing the subject categories of morbidity, lifestyle, healthcare expenditure, and healthcare outcomes,: 1) breast cancer screening rates, prevalence and incidence, 2) dementia prevalence and incidence, 3) prevalence of alcohol use by pregnant women, 4) mobility and school sports participation in children, 5) costs of smoking, 6) Dutch national healthcare expenditure, 7) hospital mortality rates, and 8) bedsores prevalence. Using selected keywords for each policy debate case, we performed a document search to identify documentation of the debates (2014-2016) on the websites of the Dutch government and parliament. We retrieved 163 documents We examined the policy debate cases through a content analyses approach.

RESULTS Sources of the statistics used in policy debates were primarily government funded. We identified two distinct functions, i.e. rhetorical and managerial use of statistics. The function of the debate is rhetorical when the specific statistic is used for agenda setting or to convince the reader of the importance of a topic. The function of the debate is managerial when statistics determine planning, monitoring or evaluation of policy. When evaluating a specific policy, applied statistics were mostly the result of routine or standardized data collection. When policymakers use statistics for a managerial function, the policy debate mirrors terms derived from scientific debates.

CONCLUSION While statistics used for rhetorical functions do not seem to invite critical reflection, when the function of the debate is managerial, i.e. to plan, monitor or evaluate healthcare, their construction does receive attention. Considering the current role of statistics in rhetorical and managerial debates, there is a need to be cautious of too much leniency towards the technocratic process in exchange for the democratic debate.

INTRODUCTION

There is an ongoing debate about the dividing line between facts and opinions, and the role of scientific knowledge in policymaking (1). Terms such as ‘fact-free politics’ (2) ‘science as an opinion’ (3) and, more recently, ‘alternative facts’ (4) reveal a concern for the position and credibility of facts in both politics and policymaking.

Although the use of evidence in policymaking remains controversial, recent developments in the political landscape in Europe and the USA have fuelled a growing concern among scientists and others who advocate the use of evidence in policy (5); this is illustrated by the March for Science movement (6), the ‘Sense About Science’ campaign (7), and the Alliance for Useful Evidence network (8). The discussion on the use of facts applies to all fields of policy but, in particular, to the field of health and healthcare, which has long been bound to the tradition of evidence-based practice and policy (9).

In the Netherlands, healthcare policy aims to support the delivery of high quality, accessible and affordable healthcare services to improve the health of the Dutch population (10). As in clinical practice, healthcare policy is increasingly expected to be based on evidence (11-13). Evidence can be used for different functions within the policy cycle, e.g. agenda setting (where evidence is used to underpin the need for policy); policy formulation (where evidence is used as a basis for policy development); implementation (where evidence is used to determine how policy can best be materialized); and monitoring and evaluation (where the (un)intended effects of implemented policies are measured, informing the need for improvement of policy and practice) (14, 15). However, policymaking is not a cyclical process fluently flowing from evidence to application and to full implementation. It is iterative and context bound, involving the consideration of many values of which evidence is one, besides ideology, practicability, the complexity of the subject, timeliness, and the distribution of power in politics (16-17).

Statistics (quantitative information) on health and healthcare constitute an important base of evidence for health policy (19-21). The construction of statistics requires social and intellectual investment that is often taken for granted (22). The government has made large investments in the development and maintenance of a data infrastructure comprising registries, survey research, and the development of statistics resulting from these various data sources (23). Nevertheless, the users of statistics may show little interest in how the statistics were constructed and/or how the underlying data were collected (22).

Espeland et al. (2008), describe how statistics foster cooperation and control in complex systems. Statistics enable policymakers to evaluate healthcare and enforce sanctions or incentives, since statistics are, seemingly, easy to interpret. If they carry authority, statistics can be used 'to persuade'. However, that authority depends on trust in the statistics' accuracy and validity, their usefulness in solving problems, how they link those who use the statistics and those who have invested in their development, and how statistics are considered to be objective, as opposed to human judgment (22).

Consequently, to understand how statistics on health and healthcare are used in the national policymaking process in the Netherlands, insight is needed into how the purpose for which statistics are used are connected with the function of the debate, and with the sources and construction of statistics in policy debates on health and healthcare (24-26). Such insight should increase awareness among researchers on how their research, as expressed in statistics, is used in policy debates in government and parliament. We therefore explore how different types of statistics are used in various ongoing policy debates on health and healthcare. In the Netherlands, most of the policy debates in government and parliament are documented and published (in written text). This allows systematic analysis of the use of statistics in parliamentary healthcare debates.

METHODS

Sampling

Based on analysis of literature and our knowledge of ongoing policy debates in health and healthcare in the Netherlands. we focused on four categories of statistics: 1) morbidity statistics, 2) lifestyle statistics, 3) healthcare expenditure data, and 4) statistics on healthcare outcomes. For each category, the use of statistics was analysed in two policy debates on different topics (Table 1).

In the Netherlands, although an important part of policymaking takes place at the municipality level, the present study focused solely on the national policy debate. This policy debate on health and healthcare is understood as the formal communications between government and parliament. All communication between government and parliament is documented and made public on their respective websites, including the minutes of parliamentary debate. These texts reflect the policy debate and are part of the policy context (27-30). Consequently, with a considerable part of the policy debate on health and healthcare being published, analysis of these documents provides insight into how statistics are used in the policy process.

For each of the four categories of statistics, through purposeful sampling, we selected two policy debate cases that encompassed the formal ongoing discussions on a health topic over a two-year period (2014-2016) (31). The aim was to include policy debates on statistics that used different methods of data collection to underpin statistics and were subject to current policy debates at the national level. To minimize researcher bias in the selection of cases, the policy debate cases on the use of statistics in policy debates were reviewed, discussed and agreed upon by all authors.

Table 1 | Policy debate cases selected for the present study.

Category	Case 1	Case 2
Morbidity statistics	Breast cancer screening rates, prevalence and incidence	Dementia prevalence and incidence
Lifestyle statistics	Prevalence of alcohol use by pregnant women	Mobility and school sports participation in children
Healthcare expenditure data	Costs of smoking	Dutch national healthcare expenditure
Statistics on healthcare outcomes	Hospital mortality rates	Bedsores prevalence

Eight cases were chosen to represent the categories morbidity, lifestyle, healthcare expenditure, and healthcare outcomes (Table 1).

Morbidity: Breast cancer screening rates, prevalence and incidence were selected because these statistics are derived from routine data collection through a national cancer registry. Dementia prevalence and incidence was selected because this figure is not measured through standardized data collection but constructed through modelling.

Lifestyle: Prevalence of alcohol use among pregnant women was included because the statistic is derived from a single published study. Mobility rates and school sports participation in children were chosen as these statistics can be derived from multiple sources.

Healthcare expenditure: Dutch national healthcare expenditure was selected because of its standardized data collection method. Costs of smoking was included because it is constructed through modelling.

Healthcare outcomes: Hospital mortality rates were selected because of their clear registration and the obligation (since 1 March 2013) for Dutch hospitals to publish mortality data. Bedsores prevalence was included because of the difficulty to establish this using standardized measurement.

Data collection

We identified documents describing the use of statistics in policy debates and source-documents through i) the national government website (*www.rijksoverheid.nl*) which contains published documents from the 11 Dutch ministries (32), and ii) the website of the parliament (*www.tweedekamer.nl*) which contains all parliamentary documents, including minutes of parliamentary debates (33).

In both websites, we restricted the search period from 8 July 2014 to 8 July 2016. For each policy debate case, we carried a document search out using selected keywords. Supplementary material 1 presents details on the search methods for the policy debate documents; 163 documents were retrieved for further analysis (listed in supplementary material 2).

For this study, since the data concerned publicly available policy documents, no ethical approval was required.

Analysis

To explore how statistics on health and healthcare are used, a we used a conventional content analysis method as described by Hsieh and Shannon (2005) (34). First, we started by immersing ourselves in the data by reading through the documents. We discussed each case in the research group to increase our understanding of the context of each debate. We then selected utterances (text fragments) in the selected documents that refer to statistics on the case. Next, we coded elements of these utterances iteratively in MaxQDA. We extracted themes and subthemes from the data. By reading through the different debate cases and the iterative coding, we constructed the main categories that frame our results. After establishing the main categories, the coding process was reiterated and refined by revisiting the text, and deductively coding the full text within these categories. We recorded observations that referred to the sources of statistics, the type of statistics, construction of the statistics, and the content of the debates; and related these observations to the different functions of use.

Analyses were performed by the first author and discussed with all co-authors during all stages of each analysis.

Table 2 | Type of documents per policy debate case.

	Breast cancer	Dementia	Alcohol use by pregnant women	Mobility in children	Costs of smoking	Dutch national healthcare expenditure	Hospital Mortality	Bedsores	Total
Report	8	5	2	5	4	24	7	3	58
Parliament question	2	3	1	1		10	1		18
Plenary debate		1	1	3		3	2	1	11
Letter to/from government	5	4	2	4	1	13	7	1	37
Appendix	1	2				7		1	11
News message			1			3			4
Speech		2							2
Explanatory memorandum				1		1	1		3
Annual report				2		9	2		13
Budget				2		2	1		5
Amendment								1	1
Total documents	16	17	7	18	5	72	21	7	**163**
Utterances included in the analyses	17	29	7	23	8	42	24	8	

RESULTS

We identified 163 documents that describe and underpin the eight policy debates. Documents included research reports (n=58), transcripts of posed parliamentary questions (n=18), transcripts of plenary debates (n=11), letters to/from the government (n=37), appendixes (n=11), newspaper messages (listed as input to a debate, (n=4), speeches (n=2), explanatory memoranda (n=3), annual reports (n=13), budget texts (n=5), and an amendment text (n=1). Table 2 presents an overview of the type of documents per policy debate case. In the debates on bedsores, hospital mortality and costs of smoking, we included documents that did not contain direct use of statistics but in which the construction of the statistics was discussed. The included research reports either described the construction of the studied statistic (20 of the 58 reports), or reported the statistic more generally without mentioning the primary source or only providing a reference (e.g. in an 'introduction' section).

Characterization of analysed debate and sources of statistics

Dementia

The policy debate on dementia focuses on the recent Deltaplan Dementia. This plan involves a program stimulating interventions/research on dementia. The name 'Deltaplan' is a metaphor, referring to the major reconstruction of the Dutch Delta, indicating that the plan aims for a comprehensive change in the field of dementia care. In all the analysed documents, the prevalence figures used concerns a statistic that was calculated by extrapolating a prevalence figure from a scientific publication (dating from 1996) from a neighbourhood to country level, and to the present time. While having the same origin and referring to the same sources, the figures used in these documents ranged from 230,000 to 260,000. In one document (an answer to a parliament question) a different prevalence figure (i.e. 80,000) was used by the minister, i.e. a statistic derived from a GP registration (NIVEL zorgregistraties Eerste lijn).

We identified 17 documents in the debate on dementia.

Breast cancer

The statistics used in this policy debate were derived from the Netherlands Cancer Registry. When a statistic on breast cancer was featured in the debate, a reference was made to this registry, either through research (institutes) providing these statistics or to the website publishing the registry data. Statistics are routinely collected for this registry. The debate centres on breast cancer screening, the development of breast cancer in society, and the financing of breast cancer treatment.

We identified 16 documents in the debate on breast cancer.

Alcohol intake in pregnancy

Debate on alcohol intake in pregnancy is part of the discourse on prevention through lifestyle change. Documents indicated that both government and parliament support a change of lifestyle behaviour through policy. Statistics on alcohol use in pregnant women were derived from a study performed by two Dutch research institutes (Trimbos Institute and TNO). The statistics were based on survey research, which was repeated in 2007, 2010, and 2014.
We identified 7 documents in the debate on alcohol Intake in pregnancy.

Mobility in children

The debate on mobility in children is also part of the discourse on prevention through lifestyle change. This debate focuses on two objectives: the participation in sports by children, and sports education in schools. The debate centres on the role of government and possible policy to increase exercise in children. Statistics used in this debate are derived from three sources: the Health Behaviour in School-aged Children (HBSC) study, and the [Lifestyle Monitor] (LSM)/ [Health survey] (GE). For sports participation, we examined the statistics used in the debate on the number of hours of physical education in schools. Statistics used to indicate the overall exercise rates by children are derived from the study [Accidents and Exercise in the Netherlands] (OBiN).
We identified 18 documents in the debate on mobility in children.

Healthcare expenditure

Healthcare expenditure includes expenditure as part of the governmental budget, as a total of expenditure for different healthcare sectors, and as total expenditure development in healthcare expenditure over the years. Statistics used are provided by Statistics Netherlands (CBS). The current debate focuses on the national policy to get a grip on health spending and to make the system more economically sustainable.
We identified 72 documents in the debate on health care expenditure.

Societal costs of smoking

Statistics on the societal costs of smoking provide a financial perspective on smoking in society. However, 'costs of smoking' does not have one commonly accepted definition. Statistics are derived from different reports/sources that either state that the societal costs of smoking are high, or that smoking does not result in increased costs for society. No statistics on the costs of smoking were used directly in the debate. To focus attention on the harmful effects of smoking, in a few supporting reports the high costs of smoking for society were mentioned, but without substantiating the argument with statistics.
We identified 5 documents in the debate on societal costs of smoking.

Bedsores

The current debate on the rate of bedsores centres on the high prevalence of bedsores in the Netherlands compared to other countries in Europe, highlighting the need to decrease this rate. All statistics used in this debate are derived from the National Prevalence Measurement of Care problems (LPZ) 2013.
We identified 7 documents in the debate on bedsores.

Hospital mortality

The current debate on hospital mortality rates focuses on the development of a standardized method of data reporting. Hospitals are obliged to publish these statistics, with the Hospital Standardized Mortality Rate (HSMR) as the intended measure. Currently, the studied debate focuses on the construction of these statistics.
We identified 21 documents in the debate on hospital mortality.

Sources across cases: Government and non-government related

A reference was frequently provided for the statistics that were used (99 of 174 statistics). These sources concerned:

- reports (n=65) (including reports that were part of our sample, n=20)
- websites (n=16)
- scientific studies (n=7)
- news articles (n=2)
- and others (n=9)

Of the 99 referenced sources, 74 concerned governmental agencies (e.g. RIVM, SCP, CBS) or organizations funded by the government to conduct research on the respective topic.

- governmental agencies (n=41)
- consultancies (n=16)
- research institutes (n=17)

Type of statistics across cases, and use of tables and visualizations

The identified datasets used both rounded (n=86) and exact statistics (n=53). Rounded statistics are those that can be rounded up or down, mentioned in words, or expressed through normative words (e.g. ‘strongly decreased’). Exact statistics are precise to a single digit in absolute form, or expressed in exact percentages. In all debates, statistics were used as rounded statistics at least once. In the debate on dementia, the prevalence figure was presented only as a rounded figure, and always including terms such as ‘more than’ or ‘approximately’.

Visualization was used to present statistics in the debates on healthcare expenditure (n=7), mobility in children (n=7), breast cancer (n=2), hospital mortality (n=5), and bedsores (n=1). These visualizations were used in reports. Also, tables were frequently used to present statistics on healthcare expenditure (n=13).

Example of a rounded figure:

Healthcare expenditure: Healthcare expenditure is the largest government expenditure after social security. In the Netherlands, we collectively spent around €71 billion on healthcare in 2015, about 11% of our GDP. (Answer to parliament question, 3 February 2015)

Example of an exact figure

Healthcare expenditure: *In 2014, collective healthcare expenditure rose by 0.1 percentage points, to 10.0% of the GDP. In 2015, the collective healthcare expenditure will decrease by 0.1 percentage points to 9.9%, equal to the level of 2012 and 2013. The nominal growth of healthcare expenditure will decrease from 2½ % in 2014 to 1¼ % in 2015. (Report: Macro economische verkenning 2015, by CPB, 10 September 2014)*

Functions of policy debates: rhetorical and managerial

We identified two functions for the use of statistics in the policy debates, i.e. rhetorical and managerial. First, when statistics are used to convince the listener to act, the function of the debate is rhetorical; the actual number does not affect its use in the debate, but seems to indicate 'a lot'. Second, statistics are used for a managerial function when planning, monitoring, or evaluating specific healthcare policy. When used managerially, the number itself is instrumental to the decisions made in planning, monitoring, or evaluating specific policy; when the number changes, so does the decision to be taken.

In the studied policy debates, when statistics were used for a managerial function, the way in which the statistics were constructed became a topic of discussion. In addition, when statistics were used to evaluate the effectiveness of a particular policy, the policymakers adopted the terms used by the research community.

Use of statistics for rhetorical functions

When used rhetorically, policymakers used statistics as argumentation tools to recruit support and to place or maintain issues on the policy agenda. When used rhetorically in the studied documents, statistics were mostly rounded and the exact number did not seem to be relevant to the discussion.

For example, in the debate on dementia, the prevalence figure was used to emphasize the problem of dementia in society.

Dementia: *"Care for people with dementia is high on the societal and political agenda. Research indicates that the number of people with dementia will increase sharply in the coming years. It is expected that by 2040 half a million Dutch people will have a form of dementia. At the moment that is half [of half a million]." (Report: 'Kijken met andere ogen naar zorg voor mensen met dementia en onbegrepen gedrag, by the Ministry of Health Welfare and Sports, June 2015)*

Another example is the debate on prevalence of breast cancer. In the following quote from a report on the reimbursement of cancer treatment, the size of the problem breast cancer in society was illustrated through a rounded statistic. Only after this rhetorical introduction, the specifics of the reimbursement of the treatment were explained. No references were provided for the incidence rate of 14.000.

Breast cancer: *"Every year, 14,000 women and 100 men are diagnosed with invasive breast cancer in the Netherlands. More and more women are using the possibility of a breast reconstruction after a breast removal operation to treat breast cancer." (Report: Voorwaardelijke toelating tot het basispakket, Voortgangsrapportage, by the National Health Care Institute, 21 June 2016)*

In addition, if the statistic is used primarily to persuade, the statistic is used rhetorically. In the debates on breast cancer, healthcare expenditure, mobility in children, bedsores, and alcohol use among pregnant women, the statistics were used to evaluate policy direction.

Two examples of this:

Mobility in children: *"Too many young people exercise too little. Less than half of the Dutch youngsters does not adhere to the norm that was chosen as a baseline for policy to stay healthy and fit. (Explanatory memorandum to a proposed bill, 25 February 2016)"*

Alcohol use in pregnant women: *"How do you explain that pregnant women who drink alcohol have started drinking more? (Parliament question, 20 August 2015)"*

The minister addressed this rhetorical question with an exact, managerial, answer:

"With this correspondence, I am informing you of the manner in which the evaluation of the alcohol Licensing and Catering act will be executed. [...]. In 2007, 2010 and 2015 the TNO carried out national polls in which, amongst others, it was asked how many women used alcohol during pregnancy and breastfeeding [...]. The new statistics show a decrease in the percentage of women drinking alcohol in the first three months of pregnancy as compared to 2010 and 2007 among all education levels: from 16.5 percent in 2007, 13.8 percent in 2010, to 6.9 percent in 2015. [...] this makes me feel optimistic. (Letter from the government, 1 February 2016)."

Use of statistics for managerial functions

Statistics are used for a managerial function when the figure itself is instrumental in the decision. For instance, in the debate on mobility in children, the percentage of children that received the recommended amount of physical education was used to argue for a specific policy, i.e. a mandatory number of hours of physical education in schools. The statistic itself is what determines the decision, as the percentage of children that received physical education was considered to be too low by the opposition.

When used managerially, statistics are most likely to be exact; however, there are some exceptions to this rule. In the debate on dementia, a rounded statistic was used managerially when it was instrumental in determining the amount invested in the Deltaplan Dementia. In the debate on mobility in children, both rounded and exact statistics were used interchangeably.

Some statistics are intended to be used managerially by the policymaker, such as the bedsores statistics and hospital mortality rate. However, in the studied documents, the application of these statistics for practical decisions was rejected. In an exchange between government and parliament, a member of parliament proposed to introduce a financial reward for those who keep the bedsores rates below a certain benchmark; however, the statistics were not considered sufficiently reliable by the minister to implement this idea (see quote).

Bedsores: *For those healthcare providers who for, at most, three percent of the total number of insured to whom care is given suffer from bedsores, malnutrition, or dehydration, receive from our minister a modest bonus for the purpose of the workplace. (Amendment proposal, rejected, 15 September 2014)*

Outside government policy, the hospital mortality statistic is used for ‘internal quality improvement in hospitals’. The government stimulates improving the transparency and credibility of the statistics. In the studied texts, the managerial use of hospital mortality statistics to drive policy action was discussed. In an intended managerial use of a mortality figure, a higher risk of mortality in the weekend was used to request the government to act. Nevertheless, the debate does not prompt action, but focuses on the credibility of the figure. This discussion is illustrated in the following section.

Scientific discussion and managerial use of statistics

The need for scientific research is repeatedly mentioned in the different debates. Policymakers discussed the reliability and credibility of statistics, especially when the statistic was used managerially. In the studied debates, when used rhetorically, the construction of the statistics was not questioned.

In the debate on hospital mortality and bedsores, policymakers did not consider the statistics to be solid enough for decision-making. Here, they adopted arguments used by the scientific community (e.g. methodological criticism on the construction of the figure). The discussion on the statistic itself was illustrated by a question posed in parliament on higher hospital mortality during the weekends. The statistic was used to identify a possible healthcare problem: higher mortality in hospitals during the weekend, implying that care during the weekend might be sub-standard. The Minister of Health addressed the issue by providing an explanation on how the statistic was constructed, using argumentation provided by the scientific community on case-mix adjustment (adjusting to a differing mix of patients with regard to illness severity) to explain that statistic. Consequently, the discussion focused on the uncertainty of what the statistic indicates, rather than how to address possible problems related to the quality of hospital care during the weekends.

The quote below illustrates how a parliament member asked a question about higher hospital mortality on the weekend. The answer by the Minister refers to published research, explaining that a ‘case mix’ causes this statistic to be higher. Thus, the Minister used the uncertainty of the methodology as an argument not to address the issue at hand.

Hospital mortality: *Do you share the opinion that a 20% higher mortality risk in the weekends is so shocking that something needs to be done about this immediately? If yes, what do you propose? [Parliament member]*

From the [Monitor adverse events in Dutch hospitals 2011/2012], that EMGO/NIVEL carried out on my request [...]. A possible explanation for this - against the background of the risk of healthcare-related damage found - is that the so-called 'case mix' of patients admitted on the weekend means that these patients are, on average, sicker. [Minister of Health]. (Parliament question, 2 September 2015)

In the debate on dementia, the prevalence figure was used managerially as a basis for the evaluation of the Deltaplan Dementia. In the quote below, two statistics are compared to clarify the use of the figure. Nevertheless, as opposed to other statistics that are used managerially, this statistic is neither exact nor the result of routine data collection, but is constructed through modelling, representing an exception to our observations regarding the use of exact statistics related to the managerial use of statistics.

Dementia: *The estimation of* ***230,000 to 250,000*** *people with dementia in the Netherlands, on which the analysis of the Deltaplan Dementia is also based, is somewhat similar to international studies providing statistics for Western Europe. The RIVM bases the statistic of* ***80,000*** *on a sample from the GP registration database. The RIVM, however, mentioned that this does not provide a complete overview because it concerns a sample and also because GPs lack a complete registration of people with dementia. (Answer to parliament question by the government, 17 March 2015)*

Evaluation of specific policy through the use of statistics

In the debates on breast cancer, healthcare expenditure, mobility in children, and bedsores, statistics were used to managerially evaluate the effectiveness of the policies set by the government. We consider the use of statistics as a 'managerial evaluation' when the research was conducted with the explicit aim to evaluate a specific policy. These evaluations are intended to inform the managerial use of statistics.

For example, statistics on breast cancer incidence and prevalence are used to evaluate established programs and treatment of breast cancer. As such, the use of the screening program was evaluated and confirmed to be effective.

Breast cancer: *The conclusion is that the population screening yields considerable health benefits and the Health Council recommends continuing and further improving population screening. [...] A total of 6,975 cases of breast cancer have been detected and the detection rate has increased to 6.9 per 1000 women tested. (Appendix to a letter to parliament, 2016)*

Construction of the statistics and evaluation of specific policy

The type of data collection and evaluation of specific policy appears to be connected. All statistics used for managerial evaluation of policy were constructed through routine or standardized data collection. Statistics on national healthcare expenditure, mobility in children, and breast cancer were either constructed through standardized or routine data collection.

The statistics on dementia, costs of smoking, and alcohol use among pregnant women, were not used for specific policy evaluation. It seems that no defined method of constructing statistics on the costs of smoking was determined, and no effort was made on a policy level to strengthen these statistics. In the debate on alcohol use among women, measurements were conducted through questionnaires and, since 2007, have been repeated twice. The methodology behind these statistics is not questioned when used; however, they are not used for the evaluation of a specific policy.

In the studied debates, the discussion on the methodology behind bedsores and hospital mortality statistics was taken up in the policy debates with a managerial function. Nevertheless, the policymakers using these statistics did not consider them to be adequate to be used for implementation and evaluation.

DISCUSSION

We explored how different types of statistics are used in a variety of ongoing policy debates on health and healthcare. Statistics used were mostly derived from research directly commissioned by the government, or produced by government organizations. The main sources of the statistics were reports, websites and (occasionally) scientific studies. We distinguished two distinct functions of policy debate: rhetorical and managerial use of statistics. If the debate's function is rhetorical, statistics were used as an argumentation tool to show the success of (or need for) a certain policy, to recruit support and to place/maintain issues on the policy agenda. The debate's function is managerial when statistics were used to determine how specific policy measures are planned, monitored, or evaluated. Statistics that are used managerially are primarily exact, expressed in tables, or visualized, while rounded statistics are primarily used for rhetorical functions. When evaluating specific policy, statistics are exclusively the result of routine or standardized data collection. Furthermore, when statistics are used (or intended) for a managerial function, the debate within policy mirrors the debate on the construction of these statistics in the scientific community.

Limitations

We explored the use of statistics in the Dutch debate on health and healthcare based on eight case studies. The debate on health and healthcare consists of numerous types of statistics and topics. A large part of the included documents was connected to one case i.e. healthcare expenditure (n=72), while other cases included fewer documents e.g. alcohol use by pregnant women (n=7) and cost of smoking (n=5). While the larger number of documents within the healthcare expenditure case allowed for a broader comparison of managerial and rhetorical use of these statistics within the same context, cases with fewer documents provided vivid insights in the use of statistics in different contexts. A wider selection of cases would have resulted in a more comprehensive insight into how statistics are used across different topics. Simultaneously, an analysis of each single policy debate case and even every single text would provide a deeper understanding of how these statistics are used in their respective contexts.

The texts involved in the policy debate were derived from two websites; however, it is likely that not all documents published on these websites are represented in this study. Nevertheless, based on the single search terms used, a large part of the discourse was identified. Additionally, policy debate in the Netherlands encompasses many actors outside a governmental setting. We decided to focus on debates around the policy processes in government and parliament. Extending the analysis to include more actors would have resulted in extended insight into their influence on the of statistics in policy debate.

Our distinction between rhetorical and managerial functions is based on our interpretation of the documents and its contents. Nevertheless, these functions may overlap, e.g. when a parliament member proposes a law to make a statement (i.e. a rhetorical function) rather than to achieve a change in regulation (i.e. a managerial function).

Our results represent only those cases that were studied and cannot be generalized to all debates on healthcare involving the use of statistics in the Dutch government and parliament. Nevertheless, our results provide insight into how statistics are used in these debates and what might be expected in others.

Interpretation

Most of the sources included in the policy debate were related to the government, as also found by an earlier study on ex-ante policy studies in the Netherlands (35). This co-production of research and policy is considered to be an essential part of evidence-based policy in the Netherlands (36).

In the studied documents, policymakers used arguments in which the scientific terms 'case mix' and 'significance' were mentioned, mirroring the language used in scientific debates. By using scientific language, the policymakers and scientific institutions may gain more authority over the policy process. However, scientific language may also be a means to shift attention from societal problems to scientific/methodological problems. Then, methodological arguments could be used to discredit the evidence if it does not align with a certain political agenda. Conversely, if a statistic is trusted to be valid and accurate (22), the actual problem at hand might be discussed rather than the construction of the statistic itself.

In the debates, only a few scientific publications were used as a source. The statistics used were published in reports or derived from websites. It seems that the usability of statistics for the evaluation of policy increases when data are collected routinely or through standardized methods. To embed statistics in the managerial policy debate, it appears worthwhile to invest in routine or standardized measurements. The results of this study support Cairney and Oliver (37) who proposed to reduce the uncertainty of research results, increasing the validity and reliability of statistics to encourage the managerial use of evidence in policy.

Further research on evidence-based policy could focus on the use of statistics in policy debate in other countries. It would be interesting to establish whether the connection between the managerial use of statistics, scientific discussion and the routine data collection of statistics is similar to those in the debates we studied in the Netherlands. Moreover, the relation between the current debates and change in debates over different time periods might provide useful insight into the managerial and rhetorical use of statistics over time. Additionally, to improve understanding of the role of managerial and rhetorical use of statistics in the decision-making process, future research could address how these functions are applied to guide choices between particular policy interventions.

Implications for policy and practice

It has been argued that policymakers lack the knowledge and time to critically evaluate the statistics they apply (37, 39). Statistics might be used in contexts and for purposes other than that for which they were initially created (40-42). The results of this study indicate that policymakers do take the time to understand the construction of statistics and refer to scientific discussions when the statistic is intended to be used managerially. Questions are asked on the statistics used to shape policy; therefore, policymakers might need to relate the construction of the statistics to defend their policy choices and might need to prepare for that discussion. While this attention to the construction of statistics seems positive for evidence-based policy, the use of statistics to lend objectivity to policy decisions has a shadow side. Currently, research is focused on 'getting more evidence into policy', rather than on the suitability of evidence in the policy process (43, 44). If healthcare policy is increasingly based on research alone, the debate on what policy may be implemented would be led by the scientific community rather than the judgment of policymakers. Consequently, policy decision-making may become a purely technocratic process, rather than a democratic one (45).

Moreover, the number of government-related sources used indicates that the statistics applied in the policy are mainly government driven. While this co-production of evidence is considered a strong feature of Dutch policymaking and could support effective implementation of evidence in policy (45), government influence might also affect the outcomes of research. Researchers should be aware that the statistics used by policymakers are primarily derived from government-related institutions and routine or standardized data collection. As the government may invest in research on topics that they are interested in, those topics that do not have government priority might become under-investigated (46). Moreover, reports funded by the government could be skewed to display a positive view of government action or policy (35, 47). Consequently, with much of the evidence used in policy being government related, researchers need to continue to be transparent concerning their methods and the nature of government involvement.

CONCLUSION

Our results indicate a rational process of integration of statistics as evidence in the policy process when used for decision-making. Whereas the statistics used for rhetorical functions do not seem to invite critical reflection, when the goal of the debate managerial, i.e. to plan, monitor or evaluate healthcare, the construction of the statistics receives consideration by parliament. Considering the current role of statistics as a rhetorical and managerial argumentation tool, there is a need to be cautious of too much reliance on statistics for all policy decisions in exchange for a balanced democratic debate in evidence-based policymaking.

REFERENCES

1. Strassheim H, Kettunen P. When does evidence-based policy turn into policy-based evidence? Configurations, contexts and mechanisms. Evidence & Policy: A Journal of Research, Debate and Practice. 2014;10(2):259-77.
2. Veldheer Vr. [The consequenses of 'fact-free' politics]. De gevolgen van 'fact-free politics' voor beleid en onderzoek. The Hague / Nijmegen: Sociaal en Cultureel Planbureau; 2012.
3. Groen TP. Startdocument 'Systeemfalen' The Hague; 2013.
4. Speed E, Mannion R. The rise of post-truth populism in pluralist liberal democracies: challenges for health policy. Int J Health Policy Manag. 2017;6(5):249-51.
5. Merino JG, Jha A, Loder E, Abbasi K. Standing up for science in the era of Trump. BMJ. 2017;356.
6. March for Science 2017 [Available from: https://satellites.marchforscience.com/.]
7. Sense about Science 2019 [Available from: https://senseaboutscience.org/.]
8. The Alliance for Useful Evidence 2019 [Available from: https://www.alliance4usefulevidence.org/about-us/aboutus/.]
9. McKee M, Greer SL, Stuckler D. What will Donald Trump's presidency mean for health? A scorecard. The Lancet. DOI: https://doi.org/10.1016/S0140-6736(17)30122-8.
10. Niessen LW, Grijseels EWM, Rutten FFH. The evidence-based approach in health policy and health care delivery. Social Science & Medicine. 2000;51(6):859-69.
11. Walshe K, Rundall TG. Evidence-based management: from theory to practice in health care. Milbank Q. 2001;79(3):429-57, IV-V.
12. Lavis J, Davies H, Gruen R, Walshe K, Farquhar C. Working Within and Beyond the Cochrane Collaboration to Make Systematic Reviews More Useful to Healthcare Managers and Policy Makers. Healthcare Policy. 2005;1(2):21-33.
13. Black N. Evidence based policy: proceed with care. BMJ. 2001;323(7307):275-9.
14. Walt G, Shiffman J, Schneider H, Murray SF, Brugha R, Gilson L. 'Doing' health policy analysis: methodological and conceptual reflections and challenges. Health Policy and Planning. 2008;23(5):308-17.
15. Onwujekwe O, Uguru N, Russo G, Etiaba E, Mbachu C, Mirzoev T, et al. Role and use of evidence in policymaking: an analysis of case studies from the health sector in Nigeria. Health Res Policy Sy. 2015;13(1):46.
16. Davis P, Howden-Chapman P. Translating research findings into health policy. Social Science & Medicine. 1996;43(5):865-72.
17. Weiss CH. The Interface between Evaluation and Public Policy. Evaluation. 1999;5(4):468-86.
18. Dobrow MJ, Goel V, Upshur REG. Evidence-based Health Policy: Context and Utilization. Soc Sci Med. 2004;58(1):207-17.
19. Rice DP. The Role of Statistics in the Development of Health Care Policy. The American Statistician. 1977;31(3):101-6.
20. AbouZahr C, Adjei S, Kanchanachitra C. From data to policy: good practices and cautionary tales. The Lancet. 2007;369(9566):1039-46.
21. Upshur RE, VanDenKerkhof EG, Goel V. Meaning and measurement: an inclusive model of evidence in health care. Journal of evaluation in clinical practice. 2001;7(2):91-6.
22. Espeland WN, Stevens ML. A Sociology of Quantification. Arch Eur Sociol. 2008;49(3):397-432.

23. Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, Gülmezoglu AM, et al. How to increase value and reduce waste when research priorities are set. The Lancet. 2014;383(9912):156-65.
24. Greenhalgh T, Russell J. Evidence-based policymaking: a critique. Perspect Biol Med. 2009;52(2):304-18.
25. Russell J, Greenhalgh T, Byrne E, McDonnell J. Recognizing rhetoric in health care policy analysis. J Health Serv Res Policy. 2008;13(1):40-6.
26. Murray SJ, Holmes D, Rail G. On the constitution and status of 'evidence' in the health sciences. Journal of Research in Nursing. 2008;13(4):272-80.
27. Vogel JP, Oxman AD, Glenton C, Rosenbaum S, Lewin S, Gulmezoglu AM, et al. Policymakers' and other stakeholders' perceptions of key considerations for health system decisions and the presentation of evidence to inform those considerations: an international survey. Health Res Policy Sy. 2013;11:19. DOI:10.1186/1478-4505-11-19.
28. Schwitzer G, Mudur G, Henry D, Wilson A, Goozner M, Simbra M, et al. What are the roles and responsibilities of the media in disseminating health information? PLoS Med. 2005;2(7):e215.
29. Young ME, Norman GR, Humphreys KR. Medicine in the Popular Press: The Influence of the Media on Perceptions of Disease. Plos One. 2008;3(10) e3552. DOI:10.1371/journal.pone.0003552
30. de Goede J, van Bon-Martens MJH, Mathijssen JJP, Putters K, van Oers HAM. Looking for interaction: quantitative measurement of research utilization by Dutch local health officials. Health Res Policy Sy. 2012;10(1):9.
31. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research. Adm Policy Ment Health. 2015;42(5):533-44.
32. Rijksoverheid. Over rijksoverheid.nl. 2017 [Available from: https://www.rijksoverheid.nl/over-rijksoverheid-nl]
33. Tweede Kamer.Tweedekamer.nl. 2017 [Available from: https://www.tweedekamer.nl/applicaties/colofon.]
34. Hsieh H-F, Shannon SE. Three Approaches to Qualitative Content Analysis. Qualitative Health Research. 2005;15(9):1277-88.
35. Haarhuis CMK, Smit M. Ex ante research explored: numbers, types and use of ex ante policy studies by the Dutch government. Evidence & Policy: A Journal of Research, Debate and Practice. 2017;13(4):647-68.
36. Bekker M, van Egmond S, Wehrens R, Putters K, Bal R. Linking research and policy in dutch healthcare: infrastructure, innovations and impacts. Evidence & Policy: A Journal of Research, Debate and Practice. 2010;6(2):237-53.
37. Cairney P, Oliver K. Evidence-based policymaking is not like evidence-based medicine, so how far should you go to bridge the divide between evidence and policy? Health Res Policy Sy. 2017;15(1):35.
38. Andermann A, Pang T, Newton JN, Davis A, Panisset U. Evidence for Health II: Overcoming barriers to using evidence in policy and practice. Health Res Policy Sy. 2016;14:17.
39. O'Donoughue Jenkins L, Kelly PM, Cherbuin N, Anstey KJ. Evaluating and Using Observational Evidence: The Contrasting Views of Policy Makers and Epidemiologists. Front Public Health. 2016;4:267.
40. Woloshin S, Schwartz LM, Casella SL, Kennedy AT, Larson RJ. Press releases by academic medical centers: not so academic? Ann Intern Med. 2009;150(9):613-8.
41. Bodemer N, Muller SM, Okan Y, Garcia-Retamero R, Neumeyer-Gromen A. Do the media provide transparent health information? A cross-cultural comparison of public information about the HPV vaccine. Vaccine. 2012;30(25):3747-56.

42. Caulfield T. The Commercialisation of Medical and Scientific Reporting. PLoS Med. 2005;1(3):e38.
43. Hawkins B, Parkhurst J. The 'good governance' of evidence in health policy. Evidence & Policy: A Journal of Research, Debate and Practice. 2016;12(4):575-92.
44. Oliver K, Lorenc T, Innvær S. New directions in evidence-based policy research: a critical analysis of the literature. Health Res Policy Sy. 2014;12(1):34.
45. Beckett K, Farr M, Kothari A, Wye L, le May A. Embracing complexity and uncertainty to create impact: exploring the processes and transformative potential of co-produced research through development of a social impact model. Health Res Policy Syst. 2018;16(1):118.
46. Goldfarb B. The effect of government contracting on academic research: Does the source of funding affect scientific output? Res Policy. 2008;37(1):41-58.
47. Boswell C, Smith K. Rethinking policy 'impact': four models of research-policy relations. Palgrave Communications. 2017;3(1). DOI: 10.1057/s41599-017-0042-z.

SUPPLEMENTARY MATERIAL 1

Document identification

Documents were identified through two websites: *www.rijksoverheid.nl* and www.tweedekamer.nl. The website *www.rijksoverheid.nl* uses a search function which finds only the exact words entered in the search bar, while www.tweedekamer.nl finds the stem of the words entered. Search results were assessed by title for relevance (excluding documents which had no relation to the topic). No documents were excluded based on a title alone. All other documents were downloaded and examined for the use of statistics. First, in the table of contents, chapters which referred to the topic were identified and read. Second, keywords were used to search for the topic within the documents. Search terms regarding the case were applied, using the stem of the topic (e.g. for the case 'mobility in children' a search was carried out using the terms "beweg"/ "beweeg" and "kind"). If an abbreviation was used within the text (e.g. "BK" for "breast cancer"), this term was additionally applied in the search function. The page containing the word was read, whilst looking for the use of statistics. Documents with less than 10 pages were read in their entirety. Documents in which statistics were used regarding the particular case were included. After inclusion of the documents for analysis, overlapping documents were removed. In addition, we included four reports referenced in the parliament questions, that were necessary to understand the statistics used in the debates. In total, 163 documents were included in the analysis. The number of documents identified per search is described in Table S1.1.

Table S1.1 | Document search methods

		www.rijksoverheid.nl		www.tweedekamer.nl		External source	All
Case	Search terms (in Dutch)	N	N included	N	N included	N included	Total N included
Dementia	dementie	322	12	298	12	1	17
Breast cancer	borstkanker	77	9	64	9	1	16
Alcohol use by pregnant women	zwanger alcohol zwanger drank	15 5	4	16 3	4	1	7
Mobility among children	bewegen kinderen leefstijl beweging kinderen leefstijl beweging kind leefstijl bewegen leefstijl kind	76 58 21 26	11	76	9	0	18
National health care expenditure data	zorguitgaven	182	29	180	46	0	72
Costs of smoking	roken kosten roken kost roken uitgaven roken MKBA	130 46 47 6	3	56 12 4	4	0	5
Bedsores	decubitus doorligwonden	24 6	3	16 9	4	1	7
Hospital mortality	ziekenhuis sterfte ziekenhuissterfte hospital standardized mortality ratio HSMR ziekenhuis mortaliteit	63 4 5 2 14	12	45 2 8 2 6	11	0	21

SUPPLEMENTARY MATERIAL 2

Included documents per policy debate case

Dementia

1. Minister van Volksgezondheid Welzijn en Sport. **Beantwoording kamervragen over het bericht 'Drankje tegen Alzheimer Souvenaid te snel op de markt'** 17-03-2015.
2. **Bijlage 1 Dementie** Unknown.
3. **Bijlage 2: Het Deltaplan Dementie** Unknown.
4. Staatssecretaris van Volksgezondheid Welzijn en Sport. **Brief Geestelijke gezondheidszorg Mantelzorg** 7-07-2015.
5. Staatssecretaris van Volksgezondheid Welzijn en Sport. **Brief Langdurige zorg** 26-02-2016.
6. Staatssecretaris van Volksgezondheid Welzijn en Sport. **Brief Zorg en maatschappelijke ondersteuning** 2-05-2015.
7. E.B. Birkenhager-Gilesse M. Breteler, F. van Harskamp, I. de Koning, A. Hofman. **De prevalantie van ouderen van de ziekte van Altzheimer, vasculaire dementie en dementie bij de ziekte van Parkinson; het ERGO onderzoek**
8. RIVM. **Een gezonder Nederland** 2014.
9. Verslag van een algemeen overleg. **Eerstelijnszorg** 22-09-2014.
10. Staatssecretaris van Volksgezondheid Welzijn en Sport. **Kamerbrief sameleven met dementie** 8-07-2015.
11. NIVEL. **Kennissynthese Vrijwilligershulp thuis bij mensen met dementie** Februari 2016.
12. IGZ. **Kijken met andere ogen naar de zorg voor mensen met dementie en onbegrepen gedrag** Juni 2015.
13. Vragen gesteld door de leden der Kamer. **Over het bericht «Drankje tegen Alzheimer Souvenaid te snel op de markt»** 4-02-2015.
14. **Toespraak staatssecretaris Van Rijn bij EU-conferentie over dementie** 9-05-2016.
15. **Toespraak van staatssecretaris Van Rijn bij de opening van de expositie 'Gezichten van dementie'** 10-03-2016.
16. Verslag houdende een lijst van vragen en antwoord. **Vaststelling van de begrotingsstaten van het Ministerie van Volksgezondheid, Welzijn en** 13-11-2014.
17. IGZ. **Zo houdt de inspectie de komende jaren toezicht op de verpleeg(huis)zorg** Unknown.

Breast cancer

1. Panteia. **Beleidsdoorlichting ziektepreventie** Maart 2015.
2. Staatssecretaris van Volksgezondheid Welzijn en Sport. **Beleidsnota Rampenbestrijding** 7-07-2016.
3. **Bijlage 1 Stand van zaken acties en activiteiten Rijksoverheid** Eind 2015.
4. Minister van Volksgezondheid Welzijn en Sport. **Brief van de minister van volksgezondheid, welzijn en sport over Geneesmiddelenbeleid** 7-04-2016.
5. Minister van Volksgezondheid Welzijn en Sport. **Brief van de minister van volksgezondheid, welzijn en sport over Herziening Zorgstelsel** 26-05-2015.
6. ZI. **Brief: Rapport Zinnige Zorg - Systematische Analyse Nieuwvormingen** 16-04-2016.
7. Trimbos instituut. **Depressiepreventie: gerichte aanpak voor risicogroepen** 2014.

8. IBMG Erasmus Universiteit Rotterdam;. **Groot onderhoud van de diagnosekostengroepen (DKG's) in het risicovereveningsmodel voor de somatische zorgkosten** 16-04-2015.
9. IGZ. **Het resultaat telt ziekenhuizen 2013** Februari 2015.
10. IGZ. **Het resultaat telt ziekenhuizen 2014** Januari 2016.
11. Vragen gesteld door de leden der Kamer. **Hormoon verstorende stoffen en de relatie met de volksgezondheid** 08-03-2016.
12. met de daarop door de regering gegeven antwoorden Vragen gesteld door de leden der Kamer. **Hormoon verstorende stoffen en de relatie met de volksgezondheid** 20-04-2016.
13. GGD West Brabant. **Incidentie van kanker in de gemeente Moerdijk (2004-2013)** Juni 2016.
14. ZI. **Pakketcriteria pertuzumab** 20-01-2016.
15. ZI. **Rapport Zinnige Zorg - Systematische Analyse Nieuwvormingen** 16-04-2015.
16. ZI. **Voorwaardelijke toelating tot het basispakket Voortgangsrapportage 2016** 21-06-2016.

Alcohol use by pregnant women

1. TNO. **Alcoholgebruik tijdens zwangerschap en borstvoeding** Publication date unknown.
2. Erasmus MC. **Eindrapportage Aanpak babysterfte in Nederland** Maart 2016.
3. Staatssecretaris van Volksgezondheid Welzijn en Sport. **Kamerbrief over voortgang alcohol januari 2016** 01-01-2016.
4. Staatssecretaris van Volksgezondheid Welzijn en Sport. **Vragen gesteld door de leden der Kamer, met de daarop door de regering gegeven antwoorden [bericht: Zwangere vrouw drinkt meer alcohol]** 28-09-2015.
5. Minister van Volksgezondheid Welzijn en Sport. **Zorg rond zwangerschap en geboorte** 07-03-2016.
6. Verslag van een algemeen overleg. **Zorg rond zwangerschap en geboorte.** 22-01-2015.
7. Nederlands Dagblad. **Zwangere vrouw drinkt meer alcohol** 19-08-2015.

Mobility among children

1. **Beantwoording begrotingsvragen Jeugd en Sport** Date unknown.
2. Verslag van een algemeen overleg. **Bewegingsonderwijs** 24-09-2014.
3. Minister van volksgezondheid welzijn en sport. **Brief Geannoteerde agenda van de Sportraad op 25 november 2014** 18-11-2014.
4. Minister van Volksgezondheid Welzijn en Sport. **Brief van de minister van volksgezondheid, welzijn en sport over Geannoteerde agenda van de Jeugd en Sportraad** 28-04-2015.
5. Staatssecretaris van Volksgezondheid Welzijn en Sport. **Commissiebrief inzake Verzoek om reactie op het HBSCrapport 2013** 09-10-2014.
6. CBS; NIVEL; NJi; NZA; RIVM; SCP; Trimbos instituut; ZIN. **De Staat van Volksgezondheid en Zorg Kerncijfers voor beleid - een introductie -** Mei, 2016.
7. **Geannoteerde agenda Sportraad 25 november 2014 te Brussel**
8. Universiteit Utrecht; Trimbos Insituut; SCP. **Gezondheid, welzijn en opvoeding van jongeren in Nederland HBSC 2013**
9. Verslag van een notaoverleg. **Initiatiefnota van het lid Rudmer Heerema: «Lichamelijke opvoeding is een vak, juist in het basisonderwijs: Een pleidooi voor beter bewegingsonderwijs op de basisschool gegeven door vakleerkrachten»** 23-05-2016.

10. **Plan van aanpak bewegingsonderwijs** 23-05-2015.
11. SCP. **Rapportage sport 2014** Januari 2015.
12. Welzijn en Sport Ministerie van Volksgezondheid. **Rijksbegroting 2015 xvi Volksgezondheid, Welzijn en Sport** 16-09-2014.
13. **Rijksjaarverslag 2014 xvi Volksgezondheid, Welzijn en Sport** 20-05-2015.
14. **Rijksjaarverslag 2015 xvi Volksgezondheid, Welzijn en Sport** 18-05-2016.
15. TNO. **Trendrapport Bewegen en Gezondheid 2000/2014** 2015.
16. **Vaststelling van de begrotingsstaten van het Ministerie van Volksgezondheid, Welzijn en Sport (XVI) voor het jaar 2016** 15-09-2015.
17. Verslag van een wetgevingsoverleg. **Vaststelling van de begrotingsstaten van het Ministerie van Volksgezondheid, Welzijn en Sport (XVI) voor het jaar 2016** 22-12-2015.
18. Memorie van toelichting. **Voorstel van wet van het lid Van Nispen tot wijziging van de Wet op het primair onderwijs, de Wet op de expertisecentra en de Wet primair onderwijs BES ter invoering van regels over de kwalificatie van docenten en het vaststellen van een minimum aantal uren voor wat betreft het bewegingsonderwijs** vergaderjaar 2015–2016.

National health care expenditure data

1. Algemene rekenkamer. **Achtergronddocument Bezuiniging op de zorgtoeslag Realisatie en effect** Unknown.
2. BPSEconomics. **Allocatie professionals in de curatieve zorg** Mei 2015.
3. **Antwoorden op Kamervragen Miljoenennota 2015**
4. Minister van Volksgezondheid Welzijn en Sport. **Antwoorden op kamervragen over hoe de kosten van het huidige zorgstelsel ongezien naar de burger verschoven** 9-02-2016.
5. CAOP. **Arbeid in Zorg en Welzijn, Jeugdzorg en Kinderopvang 2015**
6. **Beantwoording kamervragen over Eerste suppletoire begroting 2016**
7. **Beantwoording vragen aan de regering over Financieel Jaarverslag Rijk 2014** Unknown.
8. rijksoverheid. **Beleidsdoorlichting begrotingsbeleid** Juli 2016.
9. **Bijlage 2: Verticale toelichting 2016**
10. **Bijlage 2: Verticale Toelichting Najaarsnota 2014**
11. **Bijlage 2: Verticale Toelichting Najaarsnota 2015**
12. **Bijlage bij Het Nederlandse Stabiliteitsprogramma 2016**
13. Staatssecretaris van Volksgezondheid Welzijn en Sport. **Brief Arbeidsmarktbeleid en opleidingen zorgsector** 9-02-2016.
14. Minister van Volksgezondheid Welzijn en Sport. **Brief instelling van een Technische werkgroep Beheersinstrumentarium Zorguitgaven** 8-07-2016.
15. Staatssecretaris van Volksgezondheid Welzijn en Sport. **Brief Kwaliteit van Zorg** 19-02-2016.
16. Minister van Financien. **Brief Najaarsnota 2015** 27-11-2015.
17. Minister van financien. **Brief Nederlandse Stabiliteitsprogramma 2016** 12-04-2016.
18. Minister van Financien. **Brief Nota over de toestand van 's Rijks Financiën** 1-07-2016.
19. Minister van Volksgezondheid Welzijn en Sport. **Brief Nota over de toestand van 's Rijks Financiën** 6-07-2016.
20. Minister van Volksgezondheid Welzijn en Sport. **Brief Toekomstig sportbeleid** 17-09-2015.

21. Algemene rekenkamer. **Brief Toepassing van de Wet openbaarheid van bestuur** 24-03-2016.
22. Minister van Volksgezondheid Welzijn en Sport. **Brief Totstandkoming rekenpremie** 18-11-2014.
23. De Algemene Rekenkamer. **Brief Vaststelling van de begrotingsstaten van het Ministerie van Volksgezondheid, Welzijn en Sport (XVI) voor het jaar 2015** 5-11-2015.
24. Minister van financien. **Brief Voorjaarsnota 2016** 27-05-2016.
25. CPB. **Broos herstel Koopkracht stijgt Macro Economische Verkenning 2015** September 2014.
26. NIVEL. **De aantrekkelijkheid van werken in de zorg 2015 Cijfers en trends** 2016.
27. **Evaluatie normeringsystematiek gemeentefonds en provinciefonds 2010-2015** April 2016.
28. Andersson Elffers Felix. **Evaluatie van de Wet Marktordening Gezondheidszorg (WMG) en de Nederlandse Zorgautoriteit (NZa)** 1-09-2014.
29. KPMG. **Evaluatie Zorgverzekeringswet** September 2014.
30. Brief van de algemene rekenkamer. **Financieel Jaarverslag van het Rijk 2014** 18-05-2016.
31. Rijksoverheid. **Financieel Jaarverslag van het Rijk 2014** 20-05-2015.
32. **Financieel jaarverslag van het Rijk 2014 lijst van vragen en antwoorden** 11-06-2015.
33. Gezondheidsraad. **Hemofilie, hiv en de Wbmv** 21-04-2015.
34. Lijst van vragen en antwoorden. **Jaarverslag en slotwet Ministerie van Volksgezondheid, Welzijn en Sport 2014** 17 Juni 2015.
35. **Jaarverslag en slotwet Ministerie van Volksgezondheid, Welzijn en Sport 2015** 8-05-2016.
36. Lijst van vragen en antwoorden. **Jaarverslag en slotwet Ministerie van Volksgezondheid, Welzijn en Sport 2015** 17-06-2016.
37. Nieuwsbericht. **Kabinet investeert fors in verpleeghuiszorg** 15-09-2015.
38. Nieuwsbericht. **Kwaliteit van de zorg sterk verbeterd, zorguitgaven gestabiliseerd** 16-09-2014.
39. **Lijst van vragen en antwoorden over het Nederlandse Stabiliteits-programma2015** 21-04-2015.
40. **Miljoenennota 2015** 16-09-2014.
41. Rijksoverheid. **Miljoenennota 2016** 15-09-2015.
42. **Miljoenennota internetbijlagen**
43. Rijksoverheid. **Ministerie van Volksgezondheid, Welzijn en Sport (XVI) Rapport bij het jaarverslag** Mei 2016.
44. Directie Begrotingszaken. **Najaarsnota 2014** 21 november 2014.
45. Ministerie van financien. **Najaarsnota 2015** 27-11-2015.
46. Raad van State. **No.W06.14.0284/III/B** 11-09-2014.
47. Raad van State. **No.W06.16.0070/III/B** 8-04-2016.
48. **Nota over de toestand van 's rijks financiën** 16-09-2014.
49. **Nota over de toestand van 's rijks financiën** 15-09-2015.
50. Rijksoverheid. **Ontwerpbegrotingsplan Nederland** September 2015.
51. rijksoverheid. **Ontwerpbegrotingsplan Nederland** September 2014.
52. **Parlementaire enquête Woningcorporaties** 21 juni 2014.
53. CBS. **Persbericht CBS Tekort overheid 2014 met 2,3 procent gelijk aan jaar eerder** 26-03-2015.

54. Rijksoverheid. **Rapport Studiegroep Duurzame Groei** Juli 2016.
55. rijksoverheid. **Rapport werkgroep Zorg ten behoeve van de Studiegroep Duurzame Groei** Juli 2016.
56. Welzijn en Sport Ministerie van Volksgezondheid. **Rijksjaarverslag 2014 xvi Volksgezondheid, Welzijn en Sport** 20-05-2015.
57. rijksoverheid. **Rijksjaarverslag 2015 xvi Volksgezondheid, Welzijn en Sport** 18-05-2016.
58. Algemene rekenkamer. **Staat van de rijksverantwoording 2015 Rijksbrede resultaten verantwoordingsonderzoek** 2016.
59. **Taakopdracht Instellen Technische werkgroep Beheersinstrumentarium Zorguitgaven**
60. Algemene rekenkamer. **Trendrapport open data 2016** 24-03-2016.
61. TNO. **Trendrapport Bewegen en Gezondheid 2000/2014** 2015.
62. Verslag houdende een lijst van vragen en antwoorden. **Vaststelling van de begrotingsstaten van het Ministerie van Volksgezondheid, Welzijn en Sport (XVI) voor het jaar 2015** 13-11-2014.
63. Welzijn en Sport Ministerie van Volksgezondheid. **Vaststelling van de begrotingsstaten van het Ministerie van Volksgezondheid, Welzijn en Sport (XVI) voor het jaar 2015** 16-09-2014.
64. Memorie van toelichting. **Vaststelling van de begrotingsstaten van het Ministerie van Volksgezondheid, Welzijn en Sport (XVI) voor het jaar 2015** Unknown.
65. **Verslag van een wetgevingsoverleg** 23-06-2016.
66. Raad van State. **Voorjaarsrapportage Begrotingstoezicht 2015** 13-04-2015.
67. Raad van State. **Voorjaarsrapportage Begrotingstoezicht 2015** 13-04-2015.
68. Minister van Volksgezondheid Welzijn en Sport. **Vragen gesteld door de leden der Kamer, met de daarop door de regering gegeven antwoorden over het bericht dat het eigen risico niet meer te betalen is** 3-02-2015.
69. Minister van Volksgezondheid Welzijn en Sport. **Vragen gesteld door de leden der Kamer, met de daarop door de regering gegeven antwoorden over hoe de kosten van het huidige zorgstelsel ongezien naar de burger verschoven** 11-02-2016.
70. Verslag van een schriftelijk overleg. **Waardering van risico's bij publieke investeringsprojecten** 21-04-2016.
71. RIVM. **Zorgbalans 2014 De prestaties van de Nederlandse gezondheidszorg** 2014.
72. RIVM. **Zorgbalans 2014 Op hoofdlijnen** 2014.

Costs of smoking

1. Nuchter Kenniscentrum Leeftijdsgrenzen. **Alcohol& tabaksverkoop aan jongeren 2016** 2016.
2. Technische werkgroup zorgkeuzes in kaart. **Analyse van beleidsopties voor de zorg van tien politieke partijen>** Maart 2015.
3. Technische werkgroep Zorgkeuzes in Kaart. **Beantwoording feitelijke vragen door de Technische werkgroep Zorgkeuzes in Kaart** 17-04-2015.
4. WHO. **Global status report on alcohol and health** 2011.
5. HaskoningDHV Nederland B.V. **Risico's in perspectief Risicovergelijking** November 2014.

Bedsores

1. **Bijlage 2 Waardig leven met zorg Overzicht van initiatieven gericht op vernieuwing van de langdurige zorg**
2. Staatssecretaris van Volksgezondheid Welzijn en Sport. **Brief van de staatssecretaris van volksgezondheid, welzijn en sport over Langdurige zorg** 26-02-2016.

3. **Concept verslag Wet langdurige zorg** 11-09-2014.
4. Stichting KOH Meetpunt Kwaliteit. **De verpleegkundig specialist in de eerste lijn. Eindrapport.** Februari 2015.
5. Maastricht University. **Landelijke Prevalentiemeting Zorgproblemen Rapportage resultaten 2013** 2013.
6. Gewijzigd amendement van het lid agema ter vervanging van dat gedrukt onder nr. 101. **Regels inzake de verzekering van zorg aan mensen die zijn aangewezen op langdurige zorg (Wet langdurige zorg)** 15-09-2014.
7. RIVM. **Zorgbalans 2014** 2014.

Hospital mortality

1. Tweede Kamer. **Agenda Procedurevergadering** 22-01-2015.
2. NZA. **Beleidsbrief Marktscan medisch specialistische zorg 2014** 4-12-2014.
3. Verslag van een schriftelijk overleg. **Beleidsdoelstellingen op het gebied van Volksgezondheid, Welzijn en Sport** 22-06-2016.
4. Ministerie van Volksgezondheid Welzijn en Sport. **Betreft Commissiebrief Tweede Kamer inzake VSO inzake Resultaten van het Jaar van de transparantie (32620, nr.168)** 22-06-2016.
5. Minister van volksgezondheid welzijn en sport. **Brief initiatiefnota van de leden Anne Mulder en Michiel van Veen over patiëntveiligheid** 29-01-2014.
6. Minister van Volksgezondheid Welzijn en Sport. **Brief Invoering Diagnose Behandeling Combinaties (DBCs)** 16-12-2014.
7. Minister van Volksgezondheid Welzijn en Sport. **Brief Kwaliteit van zorg** 18-07-2014.
8. NIVEL CBS, NJi, NZA, RIVM, SCP, Trimbos instituut, ZIN. **De Staat van Volksgezondheid en Zorg Kerncijfers voor beleid - een introductie -** Mei 2016.
9. IGZ. **Het resultaat telt ziekenhuizen 2013** Februari 2015.
10. IGZ. **Het resultaat telt ziekenhuizen 2014** Januari 2016
11. Ministerie van Volksgezondheid Welzijn en Sport. **Kamerbrief Transparantie van kwaliteit in de curatieve zorg** 18-07-2014.
12. Verslag van een schriftelijk overleg. **Kwaliteitszorg** 10-10- 2015.
13. NZA. **Marktscan en beleidsbrief Medisch specialistische zorg 2014** December 2014.
14. IGZ. **Meer oog voor kwaliteit bij cardiothoracaal chirurgische centra** April 2015.
15. ZI. **Rapport Ontwikkeling algemene indicatoren** Januari 2016.
16. **Resultaten Rijksbegroting 2014**
17. Welzijn en Sport Ministerie van Volksgezondheid. **Rijksbegroting 2015 xvi Volksgezondheid, Welzijn en Sport** 16-09-2014.
18. Welzijn en Sport Ministerie van Volksgezondheid. **Rijksjaarverslag 2014 xvi Volksgezondheid, Welzijn en Sport** 20-05-2015.
19. Memorie van Toelichting. **Vaststelling van de begrotingsstaten van het Ministerie van Volksgezondheid, Welzijn en Sport (XVI) voor het jaar 2015** 2014.
20. Minister van Volksgezondheid Welzijn en Sport. **Vragen gesteld door de leden der Kamer, met de daarop door de regering gegeven antwoorden over het bericht dat het Nederlands ziekenhuis in het weekend** 2-09-2015.
21. RIVM. **Zorgbalans 2014** September 2014.

3

Improving Interpretation of Publicly Reported Statistics on Health and Healthcare: The Figure Interpretation Assessment Tool (FIAT-Health)

AUTHORS:
Reinie G. Gerrits, Dionne S. Kringos, Michael J. van den Berg, Niek S. Klazinga

Published in Health Research Policy and Systems, 2018; 16:20

ABSTRACT

BACKGROUND Policy makers, managers, scientists, patients, and the general public are daily confronted with figures on health and healthcare through public reporting in newspapers, webpages, and press releases. However, information on key characteristics of these figures necessary for their correct interpretation is often not adequately communicated, which can lead to misinterpretation and misinformed decision making. The objective of this research was to map the key characteristics relevant to the interpretation of figures on health and healthcare, and to develop a Figure Interpretation Assessment Tool-Health (FIAT-Health) through which figures on health and health care can be systematically assessed, allowing for a better interpretation of these figures.

METHODS Above mentioned key characteristics of figures on health and healthcare were identified through systematic expert consultations in the Netherlands on four topic categories of figures: on morbidity, healthcare expenditure, healthcare outcomes, and on lifestyle. The identified characteristics were used as a frame for the development of the FIAT- Health. Development of the tool and its content was supported and validated through regular review by a sounding board of potential users.

RESULTS Identified characteristics relevant for the interpretation of figures in the four categories relate to the figures' origin, credibility, expression, subject matter, population and geographical focus, time period, and underlying data collection methods. The characteristics were translated into a set of 13 dichotomous and 4-point Likert scale questions constituting the FIAT-Health, and 2 final assessment statements. Users of the FIAT-Health are provided with a summary overview of their answers to support a final assessment of the correctness of a figure, and the appropriateness of its reporting.

CONCLUSIONS FIAT-Health can support policy makers, managers, scientists, patients, and the general public to systematically assess the quality of publicly reported figures on health and healthcare. It also has the potential to support the producers of health and healthcare data in clearly communicating their data to different audiences. Future research should focus on the further validation of the tool in practice.

BACKGROUND

Every day numerous figures related to health and healthcare are reported in all kinds of sources. Policy makers, managers, scientists, patients, and the general public use these figures to guide their thinking on topics of health and healthcare (1-4). Based on these figures, inferences are made on the severity, magnitude or impact of a health issue in society (5), influencing the decision-making process of patients (6, 7), and public opinion, which is central to priority setting in health policy (8), and science (9).

Ideally, people base their decisions on the best available evidence, retrieving the figures which support their thinking directly from the source in which the figures are initially published. This, however, is often not the case (10-12). When looking for information on health and healthcare, people will obtain information from secondary sources and organisations they deem reliable (13), such as sources found through internet searches, newspapers, information leaflets of consumer organizations, television programmes, and scientific information provided by research institutes (14-17).

During the construction process of figures on health and healthcare, choices are made on definitions of what is counted and measured, which inclusion and exclusion criteria are used, and which methodology is applied, moreover, interests of involved parties may influence the results (18). Figures on health and healthcare are often reproduced and cited in reports, summaries, fact sheets, press releases and news messages. In this process, mistakes and misunderstandings may easily occur, or figures may be manipulated deliberately (19-21). Inadequate communication of the construction of these figures may result in a misreporting of estimates such as prevalence, disease severity and outcomes of research (3, 22, 23), eventually leading to wrongful interpretation of figures by readers of such publications (24). Furthermore, multiple sources of information may report contradicting figures on the same topic (25). Contradictory or unclear reporting may cause uncertainty regarding a health(care) topic (26, 27), creating a barrier for decision making (28, 29), causing avoidance of making a decision altogether, or even lead to misinformed decisions (30, 31). In addition, figures on similar topics create room for manipulation, such as applying the figure that fits best with the agenda of its users (31, 32) (e.g. politicians (33); patient advocacy of patient organisations (34); or media generating attention by publishing negative figures (35)).

Increasingly, attention is being paid to the translation of evidence into policy and practice (36). This is for instance illustrated by the development of the

AGREE instrument which assesses the quality of the process and reporting of clinical practice guidelines (37), the AIRE instrument which assesses the methodological quality of healthcare indicators and the connected reports (38), and the GATHER statement, which assesses the reporting practice of global health estimates (39). Such instruments, which seem to be actively used in practice, are aimed to give an in-depth assessment of the quality of research outcomes, and the detailed reporting of health estimates.

Methods aimed at the in-depth assessment of figures reported in scientific or extensive research publications are widely available. However, there is a lack of methods through which a practical assessment can be made of publicly reported statistics. Journalists, policy makers or interested citizens may question whether they can trust certain figures, but often do not have the time or inclination to dive into the world of statistics and research methodology. A need exists for an easy to use tool that supports users to gain insight in the key characteristics that contribute to the interpretation of a figure on health or healthcare (40-42). The objectives of this study are to (1) map the key characteristics relevant to the interpretation of figures on health and healthcare, and (2) to develop a Figure Interpretation Assessment Tool (FIAT-Health) enabling systematic assessment of publicly reported figures on health and healthcare, to improve the proper use of these figures by policy makers, managers, scientists, patients, and the general public.

METHODS

The design of the FIAT-Health was guided by a qualitative approach, relying on data derived from various forms of expert consultation in the Netherlands.

Four topic categories of figures on health and healthcare

A broad range of data is available on health and healthcare, resulting in numerous figures of varying origin. This study focusses on four categories of figures. These specific categories were selected because they differ strongly in their nature and type of data required, but have in common that they are commonly used in health policy, monitoring and healthcare planning. Moreover, these types of figures are daily found in the media by the general public. This wide range of figures covered by these categories is likely to generate insight in the most essential characteristics of figures that are important for their interpretation, and are generalizable to other figures on health and health care as well.

1) Morbidity figures:
Morbidity figures are key in determining the incidence and prevalence of diseases in a population, forming the first ideal step to estimate the need for health care (43). Figures on morbidity are typically collected through various sources, such as clinical registry data and health surveys.

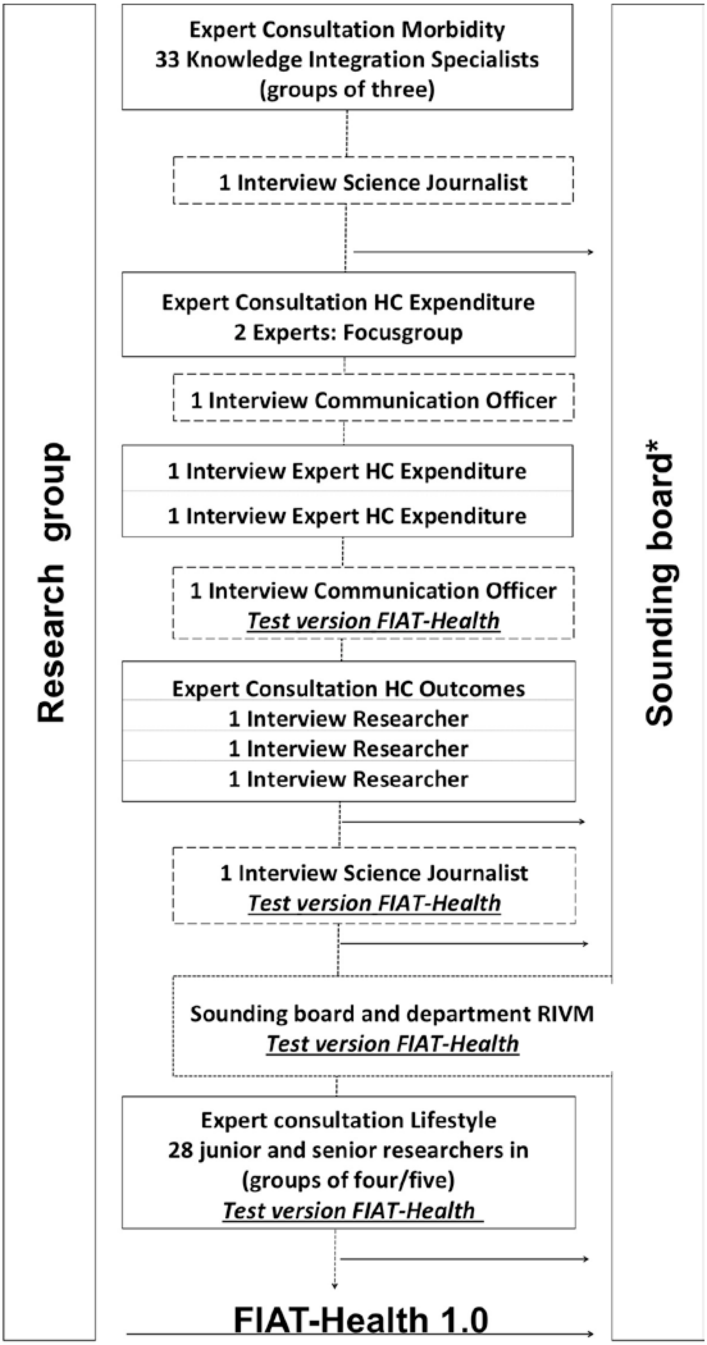
2) Healthcare expenditure figures:
Planning of health care services relies on the affordability of care, which is estimated through figures on healthcare expenditure (44, 45). Depending on the type of system, expenditure data may be based on tax data, insurance claims, providers' balance sheets, etc. cost may for example be presented in relation to the GDP, per person, or in relation to certain diseases or types of care. In this category figures are often the result of modelling.

3) Healthcare outcome figures.
The quality of healthcare services is often perceived through figures on healthcare outcomes. Through these figures the need for healthcare quality improvement and action on health is determined (46, 47). This includes clinical outcomes, such as readmissions or complications and patient-reported outcomes (PROMS). Data on healthcare outcomes is mostly derived through various clinical registries, administrative databases and patient surveys.

4) Lifestyle figures:
Lifestyle is an important determinant of health, which is needed to estimate changes in the health of a population (48). Often used examples are figures regarding physical activity, diet, and the use of alcohol, tobacco and drugs. Quantitative information on lifestyle is mostly collected using surveys and increasingly by wearable devices (49). For macro-level figures, also turnover of certain industries (e.g. tobacco) can be used.

Data collection

The research took place in the Netherlands, involving experts from four academic institutions and one national public health institute. Through purposeful sampling (50), researchers and *knowledge integration specialists* who professionally engage in the communication of research findings towards policy makers or the public, were selected, based on their extended expertise in the respective topic categories. Furthermore, science journalists and communication officers were consulted for their experience with reporting figures in the media. Expert consultations were carried out in various formats, constructed according to the emerging knowledge need, and participation of experts, as outlined below. The data collection process is described in figure 1. Data collection process. Each consultation meeting lasted between 60 to 90 minutes.



**discussion panel consisting of potential users*

Figure 1 | Data collection process

Stage 1) the first stage concerned the gathering of a broad list of characteristics on the practice of reporting figures. Data on the characteristics was obtained through the expert consultation for the morbidity category involving 33 specialists in the integration of knowledge on health and healthcare from the National Institute for Public Health and the Environment (RIVM).Fourteen groups of three experts were each asked to review the figures on a disease on four national websites from the RIVM (51-54). Some experts participated in multiple groups. Experts were asked to note the type of prevalence or incidence, the population, year, gender, age, source of the figure and the source of the data on which the figure was based. They were asked to reflect on the clarity of the figures, give their opinion on the current publication practices, identify inconsistencies that they found between the collected figures across the four websites, and identify what information was needed to improve interpretation of the figures. Their notes were compiled in an excel file and summarized.

Stage 2) as in the first stage a broad list of characteristics was derived, the second stage concerned the addition of characteristics particular to the category healthcare expenditure. A group process including discussion was considered to provide deeper insight in these characteristics. Four senior health economists from two health economy research departments, and the RIVM were invited. All four experts accepted the invitation, of which two were in the end able to attend the group session. During the session, the participants were asked to note their thoughts on the interpretation of presented figures on healthcare expenditure supported by an example of a report; consequently, the notes were structured during the session and recorded by the two present researchers. Separate appointments were made with the participants who did not attend the first session. During these individual consultations, the findings of the first meeting were discussed in-depth, complemented and validated.

Stage 3) Based on the previous stages an extended list with characteristics was developed. At this stage more in-depth questioning on the construction of figures was sought. These questions were asked during the expert consultation for the healthcare outcomes category. The consultation existed of two in-person interviews and one phone interview with three senior researchers from two academic institutions, and one private company. Open questions with a semi-structured format were used to gain insight in the considerations made both during the construction and use of figures on healthcare outcomes.

Stage 4) at the fourth stage an early version of the tool was developed, and gaps in the list of the items needed to be identified and phrasing needed to be clarified. Making use of the available expertise at the research institute, the expert consultation for the lifestyle category was addressed through a meeting with 28 junior and senior researchers of a Public Health research department. Participants received a publication on a lifestyle figure. Three groups assessed the figure through an early version of the FIAT-Health, while three other groups assessed the figure without the structured support of the draft FIAT-Health. Their findings and experiences with assessing the figure were compared and discussed in a plenary session.

Furthermore, two science journalists writing for two online journalist platforms, and two communication officers from the RIVM were individually consulted on their experience with the communication of figures on health and healthcare towards the public. Both journalists had specific knowledge on the communication of figures on health and healthcare. The consultations were aimed at the use of figures in the media, and they shared their experience on how they make figures on health and healthcare understandable to the public and policymakers.

Data analysis

Based on the expert consultations (55), insight was gained through an inductive approach in the key characteristics of each of the four topic categories of figures on health and healthcare that are relevant to their interpretation. First, to gain an overview of characteristics relevant to the interpretation of figures, characteristics related to the use of figures and characteristics related to the construction of figures were coded, resulting in an overview of characteristics for each of the four categories and successively all compiled into a single list.. Finally, characteristics in this list were synthesised resulting in the key characteristics that are essential to the interpretation of most figures on health and health and healthcare.

The format of the expert consultations differed among the categories in terms of approach and number of participants. Instead of first collecting all data in a single stage, an iterative process was applied during which multiple draft versions of the FIAT-Health were developed over time based on the emerging characteristics, and shared with the involved experts. Hence, in some sessions a preliminary draft was presented to which participants could respond, which allowed the testing of insights gained from prior sessions in subsequent consultations.

Development of the FIAT-Health

The key characteristics were structured in eight overarching themes which formed the basis of the tool development, guiding the development of the content, form, and outcome of the FIAT-Health. The themes comprised the structure on which the main components of the tool were based. Starting from this structure, the components suitable to address the themes were created.

The development of the tool was guided by constant dialogue amongst the research group, consisting of the authors of this paper. Furthermore, to gain insight in the expectations and needs of potential users of the tool, nine project leaders of the RIVM who are highly experienced in the publication of figures on health and healthcare were gathered in a sounding board. This sounding board supported the development of the tool by advising on the form, and regularly reviewing its content. The panel gathered five times, during which draft versions of the tool were reviewed in detail. Feedback resulting from these meetings was used to adapt and refine the FIAT-Health. During one additional meeting an early version of the instrument was tested by the sounding board and several knowledge integration specialists at the RIVM. In small groups they assessed an example of a published figure and reflected on the results.

Face validity was established through examination by the research group of the relevance, reasonability, unambiguousness and clarity of the content (55). Content validity was assured through the review of the tool on both content as well as form by the sounding board, thereby assuring a balanced inclusion of the aspects relevant for the interpretation of figures on health and health care (55).

RESULTS

Characteristics relevant to the interpretation of figures on health and healthcare were mapped for the four categories. Interpretation was considered to consist of characteristics relevant to the construction and use of a figure. The construction of a figure relates to the methodological considerations which impact the quality of a figure, and characteristics relevant to the use of a figure relates to the information which is needed to apply the figure in practice. As the objective was to identify the characteristics relevant to the interpretation of figures, only the characteristics related to the construction of the figure which could be understood by those without specialised knowledge were included. Furthermore, characteristics were included that were relevant to the use of a figure on health and healthcare which related to the information the user of a figure would need to assess the usability of a figure in a certain context. The resulting key characteristics that were relevant for the interpretation of each of the four categories of figures were grouped by the following eight themes:

Origin of the figure: This refers to the primary publication in which the figure was published for the first time. Any (secondary) publication that cites or refers to the figure should properly refer to this primary publication. The primary publication needs to be accessible to confirm the construction of the figure.

Credibility of the figure: Credibility is to a large extent a subjective judgement of the reader or user of a figure. This judgement is based on the expected expertise of the author and on possible conflicts of interest. Regarding expertise, a peer-reviewed scientific article from a well-known research group leads to higher expectations of credibility than an unknown blogger. Any financial or political interests may bias results or may appear to do so.

Expression of the figure: Two categories are distinguished: singular figures, such as an average or an absolute number, and composite figures, such as a percentage or a fraction. The characterisation of the figure will impact the way the figure is perceived by its reader. When a figure is communicated in a singular form, the context of that figure might be lost to its user, while a composite form could obscure the actual figure. Either way, the form will impact the way the reader will value a figure.

Subject to which the figure applies: The definition of the subject is a decision made by the authors of the publication. Often, several definitions of the seemingly same subject are available. A more broad or narrow definition can impact figures considerably. Moreover, a misconception of the definition of the subject will result in a wrongful interpretation.

Population to which the figure applies: The population forms the basis for the figure. The inclusion and exclusion criteria have an impact to whom or to what the figure applies. The interpretation of the figure needs to be supported by the knowledge of the exact population the figure applies to.

The geographical area to which the figure applies: The geographical area concerns the location to which the figure is generalizable. While the geographical area has a large impact on the figure, this characteristic is easily miscommunicated. A figure which is valid for one geographical area may not be applied to another.

The time period to which the figure applies: The time period to which the figure applies relates to the question on whether a figure is relevant to the time in which it is reported. Often, a time period to which the figure applies is not mentioned in a secondary publication, implicitly assuming that the figure applies to the present. Whether a figure of e.g. two years old is still relevant at the present time differs strongly between subjects and the aim for which it is used. Furthermore, experience reveals that figures which are counted at one point in time are often confused with longitudinal figures, i.e. point-prevalence and year-prevalence.

The process of counting and measuring: The process of counting and measuring is a broad theme covering different methodologies used to construct a figure, and the practical considerations of these methodologies. For each methodology main strengths and weaknesses need to be understood in order to interpret the figure.

Repetition of data collection: Furthermore, the repetition of data collection is considered in this theme. A figure derived from repeated or continuous data collection may be updated over time, or can be compared with earlier figures derived from the same data collection source.

Sampling: Often a sample of the population is used. A sample should be large and varied enough to represent the entire population. Statistical reliability increases with sample size and if the sample is relatively small, the reader should be careful with attaching value to it.

Registries: Registries are used as the basis for many figures on health and healthcare. Here the way the data is registered, and who registers the data is of importance. The quality of a registry, and thus the figure derived from it, depends on the completeness of data and the care with which the data is entered.

Surveys: Questions asked and answers given to a survey determine the eventual figure. Furthermore, these questions should be carefully deduced to conclusions on the subject.

Direct observations: Data can be obtained through measurements made by researchers or field workers, or in other words through direct observations. In some cases, this method has an advantage over a survey, but this may depend on the measurement instrument and the care with which the researcher measures. Obviously, many things, for instance patient experiences, cannot be observed easily.

Modelling: If a figure cannot be constructed through empirical methods, the possibility of modelling exists. In modelling, many figures are used for which assumptions are made by the modeller. These assumptions should be well grounded and transparent in order to interpret the reliability of the figure. The plausibility of certain assumptions may often be difficult to judge for many readers.

The FIAT-Health

Out of the eight themes relevant to the interpretation of figures on health and health care the Figure Interpretation Assessment Tool – Health (FIAT-Health) was constructed. The inclusion of all features was decided upon during an iterative process in which the research group thoroughly examined each item included in the tool. Five draft versions of the tool were reviewed by the sounding board. The sounding board gathered five times in order to review the emerging tool, four draft versions were presented and fine-tuned based on the received feedback. Consensus was reached on the fifth version. In the FIAT-Health a distinction is made between the primary publication, which is the source in which the figure is reported for the first time, and the publication in which the figure is assessed, which can be both a publication referring to a figure from a primary publication or the primary publication itself.

The characteristics relevant to the interpretation of figures on health and health care are addressed through a closed question format with directed routing. As multiple definitions can be applied to the terms used in the FIAT-Health, all terms and phrasing of the questions were discussed extensively and approved by both the research group and the sounding board.

The FIAT-Health exists out of thirteen main questions with sub questions, numbered 1 to 13, and two statements, numbered 14 and 15. Question 1 to 13 guide the interpretation and assessment of the figure, and consist of two types of questions: characterisation questions and assessment questions. The former are neutral, the latter express a value judgment. The characterisation questions will guide the user into understanding the figure without directing towards an assessment, while assessment questions do provide a distinction between a positive and negative answer. Three types of assessment questions are used: 1) Factual questions, which relate to the accessibility of information on a particular theme; 2) Assessment questions, for which the user is asked to give a value judgment on a particular theme; and 3) Correspondence questions, which relate to the correspondence between the primary publication and the publication in which the user is assessing the figure, which is not relevant if the figure is assessed in a primary publication.

Completing all questions results in a structured, 1-page overview of answers given to questions 1 to 13. This structured overview supports users to order their thoughts about the strengths and weaknesses of the figure. Through statement 14 and 15 users assess the correctness of the figure, and assess the appropriateness of the report of the figure respectively.

Answer format

Characterisation questions are answered using a dichotomous yes/no scale. These answers do not give a positive or negative value to the answer. Factual and correspondence questions are also answered using a dichotomous yes/no scale, on which the answer yes is continuously positive and no is negative. Answers to assessment questions are given through a numerical 4-point Likert scale (1 = negative, 4 = positive). To support the user of the FIAT-Health in assessing the figure, a 4-point scale was chosen, avoiding a middle option to stimulate users in forming an assessment (56). Concluding statements 14 and 15 are accompanied by a 5-star scale, which is easy to understand for users (57).

Furthermore, to guide the user of the FIAT-Health, the questionnaire is accompanied by a user guide including explanations of all questions.

In Table 1 the glossary used in the FIAT-Health is described. The FIAT-Health 1.0 questionnaire as provided in Table 2 was developed in Dutch. The translation to English was conducted through a forward-back translation by two bilingual translators. Discrepancies in translation were discussed within the research group and the translators until agreement was reached (55).

Table 1 | Glossary of the FIAT-Health.

Glossary	
Figure	The reported result in numbers of the process of counting and measuring.
Primary publication	The medium in which the figure was first made public, for example a report, database, website or scientific publication.
Author of the primary publication	The person/persons or organization who described the figure in the primary publication.
User	The person or organization who wants to understand, cite, or distribute the figure.
Subject	The aspect of health or care to which the figure refers, for example a disease, lifestyle factor, or treatment.
Unit	The measure in which the figure is expressed, for example persons, Euros, days or kilometres.
Population	A collection of units, which can consist of people or objects.
Process of counting	The manner in which a quantity is determined, for example, through a registration or direct observations.
Process of measuring	The manner in which the presence or size of the subject is determined through predefined values, for example, the measurement of weight based on a scale or based on a survey question.

Table 2 | The FIAT-Health 1.0 questionnaire

	Questionnaire FIAT-Health 1.0		
	Question	Answer	Routing
	What figure would you like to assess? (Provide the phrase in which the figure is mentioned.)	**[---------]**	
	Question 1. Origin of the figure		Answer question 1a.
1a.	Is the publication in which the figure is reported a primary publication?	**Yes/no**	If yes, go to question 2. If no, answer question 1b and 1c.
1b.	Is the primary publication known?	**Yes/no**	
1c.	Is the primary publication verifiable?	**Yes/no**	
	Question 2. Credibility of the figure		Answer question 2a and 2b.
2a.	How do you rate the credibility of the primary publication?	**Scale**	
2b.	How do you rate the independence of the author of the primary publication in relation to this particular figure?	**Scale**	
	Question 3. Expression of the figure		Answer question 3a.
3a.	Is the figure expressed in absolute terms?	**Yes/no**	If yes, answer question 3c. If no, answer question 3b. If the figure is not taken from a primary publication, answer question 3c.
3b.	Is the figure expressed in relative terms?	**Yes/no**	
3c.	Does the figure you are assessing match the figure in the primary publication?	**Yes/no**	
	Question 4. Subject to which the figure applies		Answer question 4a. If the figure is
4a.	How do you rate the clarity with which the subject is described in the primary publication?	**Scale**	not taken from a primary publication, answer question 4b
4b.	Does the definition of the subject of the figure you are assessing match the definition of the subject in the primary publication?	**Yes/no**	
	Question 5. Population to which the figure applies		Answer question 5a. If the figure is
5a.	How do you rate the clarity with which the population is described in the primary publication?	**Scale**	not taken from a primary publication answer question 5b.
5b.	Does the definition of the population of the figure you are assessing match the definition in the primary publication?	**Yes/no**	
	Question 6. Geographical area to which the figure applies		Answer question 6a. If the figure is not taken from a primary publication, answer question 6b.
6a.	How do you rate the clarity with which the geographical area is described in the primary publication?	**Scale**	
6b.	Does the geographical area of the figure you are assessing match the geographical area in the primary publication?	**Yes/no**	

Table 2 | Continued

	Questionnaire FIAT-Health 1.0		
	Question	Answer	Routing
	Question 7. **Time period to which the figure applies**		Answer question 7a. If the figure is not taken from a primary publication, answer question 7b.
7a.	Is the time period in which the units are counted described in the primary publication?	**Yes/no**	
7b.	Does the time period to which the figure applies match the time period in the primary publication?	**Yes/no**	
	Question 8 till 13: Methods of counting and measuring		
	Question 8. Data collection		Answer question 8a. If no, answer question 8b. If yes, go to question 9.
8a.	Are the data on which the figure is based collected periodically?	**Yes/no**	
8b.	Are the data on which the figure is based collected only once?	**Yes/no**	
	Question 9. Sample		Answer question 9a.
9a.	Is the figure based on a sample?	**Yes/no**	If yes, answer question 9b, 9c, 9d and 9e, and then question 10. If no, go to question 10.
9b.	Is the sample size known?	**Yes/no**	
9c.	Is the response known?	**Yes/no**	
9d.	Were important groups disregarded in the calculation of the figure?	**Yes/no**	
9e.	How do you rate the representativeness of the sample?	**Scale**	
	Question 10. Registration		Answer question 10a.
10a.	Were the data collected through an existing registration?	**Yes/no**	If yes, answer question 10b, if no go to question 11.
10b.	Is it known which registration was used?	**Yes/no**	If yes, answer question 10c. If no, go to question 11.
10c.	How do you rate the usability of this registration for the calculation of this specific figure?	**Scale**	
	Question 11. Survey research		Answer question 11a.
11a.	Were the data collected through survey research?	**Yes/no**	If yes, answer question 11b and 11c. If no, go to question 12.
11b.	Are the questions on which the figure is based described precisely ?	**Yes/no**	
11c.	Are the answer categories of the questions described?	**Yes/no**	If yes, answer question 11d. If no, go to question 12.
11d.	How do you rate the conclusion which was made based on the questions and the answer categories?	**Scale**	
	Question 12. Direct observations		Answer question 12a.
12a.	Are the data collected through direct observations?	**Yes/no**	If yes, answer question 12b. If no, go to question 13.

Table 2 | Continued

	Questionnaire FIAT-Health 1.0		
	Question	Answer	Routing
12b.	Is it known how the direct observations took place?	**Yes/no**	If yes, answer question 12c. If no, go to question 13.
12c.	How do you rate the accuracy of the direct observations?	**Scale**	
	Question 13. Modelling		Answer question 13a.
13a.	Was the figure constructed through modelling?	**Yes/no**	If yes, answer question 13b. If no, go to the final assessment.
13b.	Are the assumptions which were made in the model known?	**Yes/no**	If yes, answer question 13c. If no, go to the final assessment.
13c.	How do you rate the plausibility of the assumptions made in the model?	**Scale**	
	Final assessment **Your final assessment of the figure in the primary publication:**		
14.	The original figure is correct.	**[ ]**	Rate by giving 1 to 5 stars.
	Your final assessment of the publication in which the figure is reported:		
15.	The use of the figure in the report is appropriate.	**[ ]**	Rate by giving1 to 5 stars.

DISCUSSION

The purpose of this study was to develop a tool enabling a systematic assessment of publicly reported figures on health and healthcare, to improve assessment of these figures by policy makers, managers, journalists, researchers, patients, and the general public. The FIAT-Health can be used to assess publicly reported figures on health and healthcare, and is recommended to be applied by 1) policy makers to support their interpretation of figures to guide their decision making process; 2) knowledge institutes and policy advisers to provide grounded advise on the use of figures on health and healthcare to policy makers; 3) journalists, bloggers, and information officers during the writing of a public report, summary or commentary; 4) researchers for the translation of their own research to the public; and 5) patients in critically assessing figures found regarding a treatment or illness.

The tool should enable its user to 1) detect possible causes of bias associated with the construction of a figure; 2) distinguish the weaknesses and strengths associated to the methodology of a figure; 3) recognize in which context a figure may be used; 4) detect inconsistencies in the communication of the figure reported in the public and the primary publication. The FIAT-Health supports its user to systematically interpret a figure on health or healthcare, facilitating the assessment of the quality of the figure, and the assessment of the appropriateness of public report of this figure. The FIAT-Health is a practical tool avoiding extreme details while assuring coverage of all relevant aspects of the construction of the figure. Moreover, the tool should not generate a definite conclusion on the value of a figure, but should serve as a systematic guide for the thinking process of its user.

The importance of the characteristics found in this study is widely recognized to be influencing interpretation of figures, and are key to the reporting of figures on health and healthcare. Particular to this tool is the attention paid towards the clarity of definitions (58-61), and the credibility (21, 29) of the primary publication. Novel to this tool are the questions aimed to characterise the figure, which is often not addressed in any reporting guideline, but has large influence on how the figure is interpreted (62, 63). Furthermore, other themes are in accordance with the content of other tools supporting the reporting of quantitative evidence, namely the GATHER statement (39), the Drummond checklist (64), and the STROBE checklist (65). Like the GATHER statement the FIAT-Health is meant to improve both reporting as well as serve the information needs of decision-makers. The FIAT-Health is applicable to figures of varying origin. While knowledge usually relates to scientific evidence (66), the FIAT-Health is not limited to scientific evidence, and is thus usable in the many cases where figures are constructed through other means. Therefore, this tool does not dismiss the value of figures found through alternate methods, but allows the balanced interpretation of these figures.

Strengths and limitations

The main strength of the FIAT-Health is the development through collaboration with its potential users, ensuring the relevance of the content of the tool and its suitability for application in professional settings. As the tool is primarily based on the perspectives of both the research and funding institution, it is likely the language used, as well as characteristics may be biased towards their research paradigm. The involvement of other experts with a different affinity with the problem might have resulted in fewer items. In subsequent research, a larger number of institutes should be involved to broaden the perspective of the tool. Experts involved in the consultation rounds were sensitive to the problem of inadequate reporting. To avoid bias in the selection of experts, different groups of experts were involved including junior and senior researchers, and knowledge integration specialists. Additionally, journalists were consulted. Researcher bias in the identification of items was avoided through regular interaction with the research group and the sounding board. Furthermore, focus groups may incite "group think" where participants may adapt their opinion to fit the group. This bias was avoided by organizing different forms of consultation, including individual interviews during the developmental process.

The development of the FIAT-Health was supported by a sounding board, a method comparable with a nominal group technique involving the review of material received in advance and discussion amongst experts forming consensus (67). This study deviates from a nominal technique as the sounding board in this study involved several rounds and focussed on a qualitative assessment. Although the FIAT-Health is based upon the assumption that better information on evidence leads towards better decisions, it is not a decision support tool such as the SUPPORT tools (68), which have the goal to guide policy makers directly.

Fully understanding all methodological characteristics of figures, and overseeing its consequences is often extremely challenging and time consuming, even for those who work in research. Figures communicated in the media are often biased and misunderstood. The FIAT-Health is by no means a panacea that will fully solve this problem. While developing the tool, one of the main challenges was to find the right balance between thoroughness on the one hand, and practicality and compactness on the other. It may be unavoidable that the tool is too complex and time consuming for some, and too superficial for others. Consequently, the expectation that an average newspaper-reader will invest time in assessing figures using the FIAT-Health is unrealistic; neither will an experienced statistician obtain great revelations using the FIAT-Health. To aid these people, however, is not the purpose of this research. The FIAT-Health fills the critical gap between the expert and the news-consumer, existing of policy makers and advisors, journalists, managers, patients and those of the general public whose decisions depend on the correctness of a figure and its reporting.

Steps ahead

Although face and content validity are established, the FIAT-Health 1.0 is not yet tested for user experience and construct validity. Therefore, a validation study is carried out during the next stage of this study. The current paper format of the tool does not facilitate easy usage, creating a possible barrier to the uptake of the instrument. To encourage uptake of the FIAT-Health into practice, an online version is under construction, facilitating the user to be routed through the questions efficiently, after which an overview of the answers can be provided. Nevertheless, despite its current practical limitations, the questions provided in the FIAT-Health are highly relevant, and can thus be applied in every-day practice immediately. As such, a sample of the content of the RIVM website '*de Staat van Volksgezondheid en Zorg*' [51] was systematically assessed using the FIAT-Health 1.0. The implementation of the FIAT-Health in the improvement process of the website is currently being explored.

CONCLUSION

The FIAT-Health is a tool enabling systematic assessment of publicly reported figures on health and healthcare, to support a better understanding and interpretation of these figures by policy makers, managers, researchers, patients, and the general public. The use of the tool results in a 1-page overview, representing the main strengths and weaknesses of a figure. It fills the gap in scientifically developed methods which support the public reporting of figures on health and health care. As few systematic methods are available through which figures on health and health can be interpreted, the FIAT-Health adds a practical approach through which users are better informed and supported in their decision-making processes.

REFERENCES

1. Vogel JP, Oxman AD, Glenton C, Rosenbaum S, Lewin S, Gulmezoglu AM, et al. Policymakers' and other stakeholders' perceptions of key considerations for health system decisions and the presentation of evidence to inform those considerations: an international survey. Health Res Policy Sy. 2013;11.
2. Schwitzer G, Mudur G, Henry D, Wilson A, Goozner M, Simbra M, et al. What are the roles and responsibilities of the media in disseminating health information? PLoS Med. 2005;2(7):e215.
3. Young ME, Norman GR, Humphreys KR. Medicine in the popular press: the influence of the media on perceptions of disease. PLoS One. 2008;3(10):e3552.
4. de Goede J, van Bon-Martens MJH, Mathijssen JJP, Putters K, van Oers HAM. Looking for interaction: quantitative measurement of research utilization by Dutch local health officials. Health Res Policy Sy. 2012;10(1):9.
5. Williamson JML, Skinner CI, Hocken DB. Death and illness as depicted in the media. Int J Clin Pract. 2011;65(5):547-51.
6. Holmes-Rovner M, Charles S. The mammography screening controversy: who and what is heard in the press? Patient Educ Couns. 2003;51(1):75-81.
7. Passalacqua R, Caminiti C, Salvagni S, Barni S, Beretta GD, Carlini P, et al. Effects of media information on cancer patients' opinions, feelings, decision-making process and physician-patient communication. Cancer. 2004;100(5):1077-84.
8. Burstein P. The impact of public opinion on public policy: A review and an agenda. Polit Res Quart. 2003;56(1):29-40.
9. Weingart P. Science and the media. Res Policy. 1998;27(8):869-79.
10. Russell J, Greenhalgh T, Byrne E, McDonnell J. Recognizing rhetoric in health care policy analysis. J Health Serv Res Policy. 2008;13(1):40-6.
11. Greenhalgh T, Russell J. Evidence-based policymaking: a critique. Perspect Biol Med. 2009;52(2):304-18.
12. Black N. Evidence based policy: proceed with care. BMJ. 2001;323(7307):275-9.
13. Turney J. Public understanding of science. Lancet. 1996;347(9008):1087-90.
14. Dobbins M, Jack S, Thomas H, Kothari A. Public health decision-makers' informational needs and preferences for receiving research evidence. Worldviews Evid Based Nurs. 2007;4(3):156-63.
15. Ritter A. How do drug policy makers access research evidence? Int J Drug Policy. 2009;20(1):70-5.
16. Zardo P, Collie A. Type, frequency and purpose of information used to inform public health policy and program decision-making. Bmc Public Health. 2015;15.
17. McVay AB, Stamatakis KA, Jacobs JA, Tabak RG, Brownson RC. The role of researchers in disseminating evidence to public health practice settings: a cross-sectional study. Health Res Policy Syst. 2016;14(1):42.
18. Bekelman JE, Li Y, Gross CP. Scope and impact of financial conflicts of interest in biomedical research: A systematic review. JAMA. 2003;289(4):454-65.
19. Woloshin S, Schwartz LM, Casella SL, Kennedy AT, Larson RJ. Press releases by academic medical centers: not so academic? Ann Intern Med. 2009;150(9):613-8.
20. Bodemer N, Muller SM, Okan Y, Garcia-Retamero R, Neumeyer-Gromen A. Do the media provide transparent health information? A cross-cultural comparison of public information about the HPV vaccine. Vaccine. 2012;30(25):3747-56.
21. Caulfield T. The Commercialisation of Medical and Scientific Reporting. PLoS Med. 2005;1(3):e38.

22. Frost K, Frank E, Maibach E. Relative risk in the news media: A quantification of misrepresentation. Am J Public Health. 1997;87(5):842-5.
23. Yavchitz A, Boutron I, Bafeta A, Marroun I, Charles P, Mantz J, et al. Misrepresentation of Randomized Controlled Trials in Press Releases and News Coverage: A Cohort Study. PLoS Medicine. 2012;9(9):e1001308.
24. Simmerling A, Janich N. Rhetorical functions of a 'language of uncertainty' in the mass media. Public Underst Sci. 2015.
25. Goldman SR, Scardamalia M. Managing, Understanding, Applying, and Creating Knowledge in the Information Age: Next-Generation Challenges and Opportunities. Cognition Instruct. 2013;31(2):255-69.
26. Schwitzer G. Addressing tensions when popular media and evidence-based care collide. Bmc Med Inform Decis. 2013;13.
27. Andermann A, Pang T, Newton JN, Davis A, Panisset U. Evidence for Health II: Overcoming barriers to using evidence in policy and practice. Health Res Policy Sy. 2016;14:17.
28. Oliver K, Innvar S, Lorenc T, Woodman J, Thomas J. A systematic review of barriers to and facilitators of the use of evidence by policymakers. BMC Health Serv Res. 2014;14:2.
29. de Goede J, Putters K, van der Grinten T, van Oers HA. Knowledge in process? Exploring barriers between epidemiological research and local health policy development. Health Res Policy Syst. 2010;8:26.
30. Fischhoff B, Davis AL. Communicating scientific uncertainty. P Natl Acad Sci USA. 2014;111:13664-71.
31. Furedi A. The public health implications of the 1995 'pill scare'. Hum Reprod Update. 1999;5(6):621-6.
32. When Evidence Collides With Anecdote, Politics, and Emotion: Breast Cancer Screening. Annals of Internal Medicine. 2010;152(8):531-2.
33. Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S. Helping Doctors and Patients Make Sense of Health Statistics. Psychol Sci Public Interest. 2007;8(2):53-96.
34. McKeganey N, Russell C. Tobacco plain packaging: Evidence based policy or public health advocacy? Int J Drug Policy. 2015;26(6):560-8.
35. Sandell T, Sebar B, Harris N. Framing risk: communication messages in the Australian and Swedish print media surrounding the 2009 H1N1 pandemic. Scand J Public Health. 2013;41(8):860-5.
36. Grimshaw JM, Eccles MP, Lavis JN, Hill SJ, Squires JE. Knowledge translation of research findings. Implementation Science. 2012;7(1):1-17.
37. Agree Collaboration. Development and validation of an international appraisal instrument for assessing the quality of clinical practice guidelines: the AGREE project. Qual Saf Health Care. 2003;12(1):18-23.
38. J.S. dK, M. K, Klazinga NS. Prestatie-indicatoren langs de meetlat – het AIRE instrument. TVGW; 2007. p. 261.
39. Stevens GA, Alkema L, Black RE, Boerma JT, Collins GS, Ezzati M, et al. Guidelines for Accurate and Transparent Health Estimates Reporting: the GATHER statement. PLoS Med. 2016;13(6):e1002056.
40. Walker N, Bryce J, Black RE. Interpreting health statistics for policymaking: the story behind the headlines. The Lancet. 2007;369(9565):956-63.
41. Peters E. Beyond Comprehension: The Role of Numeracy in Judgments and Decisions. Curr Dir Psychol Sci. 2012;21(1):31-5.

42. Han PK, Moser RP, Klein WM, Beckjord EB, Dunlavy AC, Hesse BW. Predictors of perceived ambiguity about cancer prevention recommendations: sociodemographic factors and mass media exposures. Health Commun. 2009;24(8):764-72.
43. Jones IR. Health needs and their assessment: demography and epidemiology. In: Bowling A, editor. Research methods in health, investigating health and health services. 2 ed. Buckingham: Open University Press; 2002. p. 73.
44. Morris S, Devlin N, Parkin D, Spencer A. Economic Analysis in Healthcare: Wiley; 2012.
45. OECD, Eurostat, WHO. A System of Health Accounts. OECD Publishing; 2011.
46. Geddis PW. Health service performance indicators. Ulster Med J. 1988;57(2):121-8.
47. Porter ME. What Is Value in Health Care? New England Journal of Medicine. 2010;363(26):2477-81.
48. Mackenbach JP, Stronks K. Inleiding. In: Mackenbach JP, Stronks K, editors. Volksgezondheid en gezondheidszorg. 8 ed. Houten: Bohn Stafleu van Loghum; 2016. p. 36-7.
49. Loveday A, Sherar LB, Sanders JP, Sanderson PW, Esliger DW. Technologies That Assess the Location of Physical Activity and Sedentary Behavior: A Systematic Review. Journal of Medical Internet Research. 2015;17(8):e192.
50. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research. Adm Policy Ment Health. 2015;42(5):533-44.
51. Volksgezondheidenzorg.info. https://www.volksgezondheidenzorg.info/ 2016 [Available from: https://www.volksgezondheidenzorg.info/.
52. De Staat van Volksgezondheid en Zorg. https://www.staatvenz.nl/ 2016 [
53. Nationaalkompas.nl. http://www.nationaalkompas.nl/index.html 2016 [Available from: http://www.nationaalkompas.nl/index.html.
54. De VTV-2014. http://www.eengezondernederland.nl/Een_gezonder_Nederland 2016 [Available from: http://www.eengezondernederland.nl/Een_gezonder_Nederland.
55. Bowling A. Research methods in health, investigating heatlh and health services. 2 ed. Buckingham: Open University Press 2002.
56. Nadler JT, Weston R, Voyles EC. Stuck in the Middle: The Use and Interpretation of Mid-Points in Items on Questionnaires. The Journal of General Psychology. 2015;142(2):71-89.
57. Shwartz M, Restuccia JD, Rosen AK. Composite Measures of Health Care Provider Performance: A Description of Approaches. Milbank Quarterly. 2015;93(4):788-825.
58. Anema HA, van der Veer SN, Kievit J, Krol-Warmerdam E, Fischer C, Steyerberg E, et al. Influences of definition ambiguity on hospital performance indicator scores: examples from The Netherlands. Eur J Public Health. 2014;24(1):73-8.
59. Dorn HF. Methods of measuring incidence and prevalence of disease. Am J Public Health Nations Health. 1951;41(3):271-8.
60. Rosen T, Olsen J. Invited commentary: the art of making questionnaires better. Am J Epidemiol. 2006;164(12):1145-9.
61. Jaime Caro J, Eddy DM, Kan H, Kaltz C, Patel B, Eldessouki R, et al. Questionnaire to assess relevance and credibility of modeling studies for informing health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. Value Health. 2014;17(2):174-82.
62. Ancker JS, Kaufman D. Rethinking health numeracy: a multidisciplinary literature review. J Am Med Inform Assoc. 2007;14(6):713-21.
63. Yamagishi K. When a 12.86% mortality is more dangerous than 24.14%: implications for risk communication. Applied Cognitive Psychology. 1997;11(6):495-506.

64. Drummond MF, Sculpter MJ, Torrance GW, O'Brien BJ, Stoddart GL. Methods for the Economic Evaluation of Health Care Programmes. 3rd ed. Oxford: Oxford University Press; 2005.
65. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. Int J Surg. 2014;12(12):1495-9.
66. Contandriopoulos D, Lemire M, Denis J-L, Tremblay É. Knowledge Exchange Processes in Organizations and Policy Arenas: A Narrative Systematic Review of the Literature. The Milbank Quarterly. 2010;88(4):444-83.
67. Jones J, Hunter D. Consensus methods for medical and health services research. Bmj. 1995;311(7001):376-80.
68. Lavis JN, Oxman AD, Lewin S, Fretheim A. SUPPORT Tools for evidence-informed health Policymaking (STP). Introduction. Health Res Policy Syst. 2009;7.

4

Figure Interpretation Assessment Tool-Health (FIAT-Health) 2.0: from a scoring instrument to an assessment tool

AUTHORS:

Reinie G. Gerrits, Michael J. van den Berg, Niek S. Klazinga, Dionne S. Kringos

Published in BMC Medical Research Methodology, 2019; 19:160

ABSTRACT

BACKGROUND Statistics are frequently used in health advocacy to attract attention, but are often misinterpreted. The Figure Interpretation Assessment Tool–Health (FIAT-Health) 1.0 was developed to support systematic assessment of the interpretation of figures on health and health care. This study aimed to test and evaluate the FIAT-Health 1.0 amongst its intended user groups, and to compare the assessment of figures with the FIAT-Health to expert assessments.

METHODS Thirty-two potential users were asked to assess one publicly reported figure using the FIAT-Health 1.0, and to motivate their assessments and share their experience in using the FIAT-Health. In total four figures were assessed. For each figure, an expert on the specific topic (n=4) provided a comparative assessment. A qualitative comparative analysis of the motivations for assessment by the experts and participants was made. The consistency of the answers across participants was calculated, and answers to the evaluation questions were qualitatively analysed. Based on the results, a new version of the FIAT-Health was developed and tested by 27 employees of the National Institute for Public Health and the Environment (RIVM), and approved by the project's advisory group.

RESULTS Participants using the FIAT-Health 1.0 and experts gave similar motivations for their assessments. The motivations provided by experts aligned with the items of the FIAT-Health. Seventeen out of twenty-six dichotomous questions were consistently answered. Numerical assessment questions showed inconsistencies in how participants responded. In the evaluation, participants most frequently mentioned that thanks to its structured approach, the FIAT-Health contributed to their awareness of the main characteristics of the figure (n=14), but they did find the tool complex (n=11). The FIAT-Health 1.0 was revised from a scoring instrument into an assessment tool: the FIAT-Health 2.0, which was tested and approved by employees of the RIVM and the advisory group.

CONCLUSION The FIAT-Health 1.0 is a useful assessment tool, enabling most users to make critical and systematic assessments of figures, using similar considerations as experts. It is less useful as a quantitative scoring instrument. The FIAT-Health 2.0 can support policymakers, communication officers, and researchers to form a structured interpretation of a publicly reported figure on health(care).

BACKGROUND

Statistics on health and healthcare gain much attention in public media. Figures are being published, cited, and summarized in press releases, newsletters, and news items every day (1, 2). Moreover, in science communication, statistics are a persuasive tool for health policy advocacy (3-5). Politicians, policy makers and journalists like to use so-called “killer stats”; headline-grabbing statistics that immediately grasp the attention of a specific audience. The complex character and methodological background, necessary to really understand these figures, often gets lost in translation (6-8). Without the proper reporting of the background and methodology, figures are likely to be misinterpreted (9, 10). Misinterpretation of these figures is problematic, as they may impact policy and practice (11, 12). Spiegelhalter (2017) described the traditional information flows from statistical sources to the public (13). First, statistics developed by (A) academic and industry scientific research are reported in scientific publications, or (B), commissioned analytic and survey research statistics are reported by policy makers, official statistic bureaus, NGO's or other institutions. Second, press offices and communication departments report statistics to traditional media and online sources. Finally, through these sources the information is received by the public. In this communication flow, many questionable interpretation- and communication practices can occur, such as not reporting uncertainties, providing contexts or comparative perspectives, and providing relative but not absolute risk.

In the scientific community, many checklists and methods are available for the detailed appraisal and reporting of empirical studies, such as the EQUATOR guidelines (14). Furthermore, recently the GATHER statement (15) was published to support the reporting of findings of Global Health Estimates targeted at researchers and decision-makers. However, there is a lack of systematic methods for the reporting and appraisal of publicly reported statistics (16). Policy makers and civil society have other information needs than researchers when they interpret a figure (17, 18). While researchers often need in-depth information on the underlying statistical methods, those with less technical knowledge have few methods for the interpretation of a published figure (19).

Therefore, we developed a method for the systematic appraisal of figures on health and healthcare: The Figure Interpretation Assessment Tool – Health (FIAT-Health) (20). The FIAT-Health offers a systematic method for quantitatively assessing publicly reported figures on health and healthcare to be used by policy makers, managers, researchers, and the general public. The added value of this instrument is that its use requires little technical or methodological expertise. The first version, i.e. the FIAT-Health 1.0, consisted of 15 questions, which allow its user to better understand and interpret figures. In total 35 sub questions were included in the FIAT-Health covering factual dichotomous questions, to be answered by yes or no, assessment questions where the user assesses a characteristic of the figure on a scale from 1 to 5, and two final questions in which the user gives an overall assessment of the correctness of the figure and the appropriateness of the reporting of a figure on a scale from 1 to 4. Furthermore, a detailed explanation is provided for each question. The FIAT-Health was developed through consultation of 68 experts in four phases, and with the involvement of a sounding board (advisory group). The development of the FIAT-Health 1.0 was published elsewhere (20). Face and content validity of the tool were established during the development of the FIAT-Health (20), but its usability has not been tested amongst its intended user groups, which is fundamental to the uptake of the tool in practice (21). To further improve the usability of the FIAT-Health, the current study intends to test and evaluate the FIAT-Health 1.0 amongst its intended user groups. To find out to what extent users were able to make adequate assessments, we compared their assessments of figures with the FIAT-Health to an assessment made by experts on the specific topic who did not use the FIAT-Health.

METHODS

The study took place in the Netherlands during February – August 2017, involving participants from healthcare institutes from different regions.

Case studies

Four different publicly reported figures were selected as cases, including: the prevalence of Dutch people experiencing burnout complaints (case 1), the number of hours of intensive sports that reduces mortality risk (case 2), the financial profit from a decreasing number of Dutch smokers (case 3), and the number of premature deaths in people with dementia due to wrong medication (case 4).

The cases were selected based on a variation in primary publication, i.e. reports and peer-reviewed publications, the type of public report, and the expected quality of the publication as determined by the research group. Publications of which Amsterdam UMC, location Academic Medical Centre (AMC) and the National Institute for Public Health and Environment (RIVM) were primary authors, were not included given the affiliation of the authors.

Each participant assessed one publicly reported figure. Each figure (case) was assessed by two participants of each user group.

Data collection process

In the first stage, the FIAT-Health 1.0 was put into an Excel format, to allow for the structured use of the tool and to provide participants with a systematic overview of their answers in the intended format. Within the Excel file an evaluation form with six open-ended questions was included: *1. How do you experience the use of the FIAT-Health 1.0? 2. Which considerations had the largest impact on your evaluation regarding the correctness of the figure? 3. Which considerations had the largest impact on your assessment of reporting of the figure? 4. Did you experience any problems when using the FIAT-Health 1.0? 5. Were any important considerations missing in the FIAT-Health 1.0? 6. Do you have any suggestions for the improvement of the FIAT-Health 1.0?*

Participants and recruitment

In the second stage, participants were asked to test the FIAT-Health 1.0.

Four potential user groups were included in the study through purposeful selection: eight policy makers, eight researchers, eight communication officers, and eight students. Participants were selected from the professional network of the project team, and worked at the Ministry of Health, Welfare and Sports; the Dutch Healthcare Authority; municipalities; research institutes and universities in the Netherlands. Participating students were graduate students in medicine and public health of whom four were interns at the Amsterdam UMC, location AMC who had no professional relationship with the project team. Participants had no previous knowledge of the study.

In total 44 participants were invited and informed on the objective and methods of the study through e-mail. One policy maker, one researcher, three communication officers, and four students declined participation. Three students did not respond. In total 32 people participated in this study.

Participants who accepted the invitation received an e-mail explaining the process of participation, and they received the FIAT-Health 1.0 in Excel format including the evaluation form that participants were asked to fill in.

Furthermore, participants received the publicly reported figure (a newspaper or web publication), and the primary publication (a research report or peer-reviewed scientific publication).

Participants e-mailed their assessment and evaluation in the Excel file to RG, who collected all answers.

Expert assessment

In the third stage, to compare the assessments by the participants with the assessments by experts, four leading researchers from different universities, with a professorship in organisational psychology, sports medicine, health economics, and population health sciences respectively, were approached and asked to provide an expert assessment of one of the four figures that matched their expertise. The experts did not receive the FIAT-Health 1.0. They were asked to provide their assessment of the correctness of the figure and were asked to rate the figure with 1 to 5 stars (the last two assessment questions of the FIAT-Health) and motivate their assessment. To date, no systematic method has been used for advising policy makers on figures, who mostly ask advice from leading researchers. As an expert assessment of a figure is current practice, we considered their assessment as the "gold standard" (22) for comparison with the assessment resulting from the FIAT-Health 1.0. Furthermore, their explanations for their assessments were used to compare with the motivations by the participants.

Analyses

A qualitative comparative analysis of the motivations for assessment by the experts and the participants was made. First, from the explanation experts provided, motivations for their assessment were extracted. Second, motivations were compared and listed. Third, the participants' answers to evaluation questions 2 and 3 were coded into distinct motivations for assessment. Fourth, these motivations were categorised and compared to the expert motivations. Answers by experts and participants to the final assessment questions on the correctness of the figure and the reporting of the figure were compared.

The evaluation by the participants was derived from the answers to evaluation questions 1, 4, 5 and 6, and coded into common topics. All analyses were completed in Excel.

Moreover, to be able to see what questions may need revision, the agreement between participant answers on the numerical questions was calculated. Answers to dichotomous questions were considered inconsistent if the answer of two or more participants deviated from the majority for at least two cases. The answers given to the assessment questions were considered as inconsistent if three or more answers deviated from the majority for at least two cases.

FIAT-Health 2.0

Finally, in the fourth stage of the study, we adapted the FIAT-Health and tested the FIAT-Health version 2.0. A first revision was presented to 27 scientific staff members at the RIVM, who pilot-tested the revised FIAT-Health. Two cases of publicly reported figures were each assessed using the FIAT-Health by three groups of four or five people.

Findings and experiences with assessing the figure were discussed in a plenary session. RG made notes during the discussion, and collected the notes made during the test case by the participants. The FIAT-Health was adapted according to the feedback received. Consensus on the final version was obtained during a meeting with the sounding board involved in the development of the FIAT-Health. The English version of the FIAT-Health 1.0 was aligned with the changes made to the Dutch version by RG. The revised English version was checked and refined by a native speaker.

The process of data collection is illustrated in figure 1.

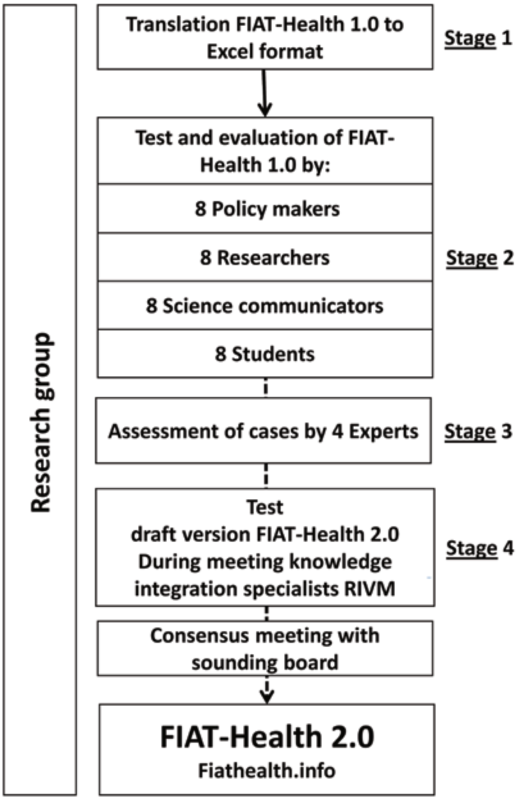


Figure 1 | Data collection process

RESULTS

The motivations provided by experts for their assessment resembled all items included in the FIAT-Health, aside from the motivation 'knowledge of the type of methodology'. Participants using the FIAT-Health 1.0 mentioned as a motivation the trustworthiness of the figure, the possibility to verify the content of the figure, and the mentioning of new information in the publicly reported message. These motivations were not mentioned by the experts. Experts used the additional motivation of knowledge of type of methodology, and their disapproval of that particular method. One participant also mentioned familiarity with that same method and rated the correctness of the figure negatively, while the participant rated the figure positively. All motivations provided by experts and participants are listed in table 1.

Table 1 | Motivations provided for the final assessment rating by experts and participants

Motivations provided by both experts and participants
• The correctness of the methods
• Assumptions on which the model is based
• Match between the primary publication and the reported figure
• Transparency on the definition of the subject
• The conclusion that was made based on the results
• Previous knowledge of the subject
• Application of the figure in practice
• An extrapolation was made
• The geographical area the figure applies to
• It concerned an estimation
• No better figures are known about the subject
• Source of the figure
• Time period to which the figure relates
• Match between the population of the reported figure and the primary publication
• Generalization of the figure
• Interpretation of the journalist
• Difference in jargon between the primary publication and the reported figure
Motivations provided by participants
• Credibility of the author
• Verifiability of the figure
• New information in the [publicly reported message]
Motivations provided by experts
• Method of modelling (the figure has no meaning as the expert considered the construct to be invalid)

A comparison between the answers by participants and the experts to the final questions on the correctness of the figure (nr. 14) and the appropriateness of the report (nr. 15) is provided in figure 2. Answers were provided on a scale from 1 (negative) to 5 (positive). Participant frequently rated both the correctness of a figure and the appropriateness of the report positively, rating 4 or 5. Experts only provided average (3) or negative (1 or 2) ratings. Participants rated the correctness of the figures higher or equal to the appropriateness of the report. Experts however, gave the same rating to the correctness of the figure and the appropriateness of its report. Only in case 4, the overall rating by participants was lower than the expert rating.

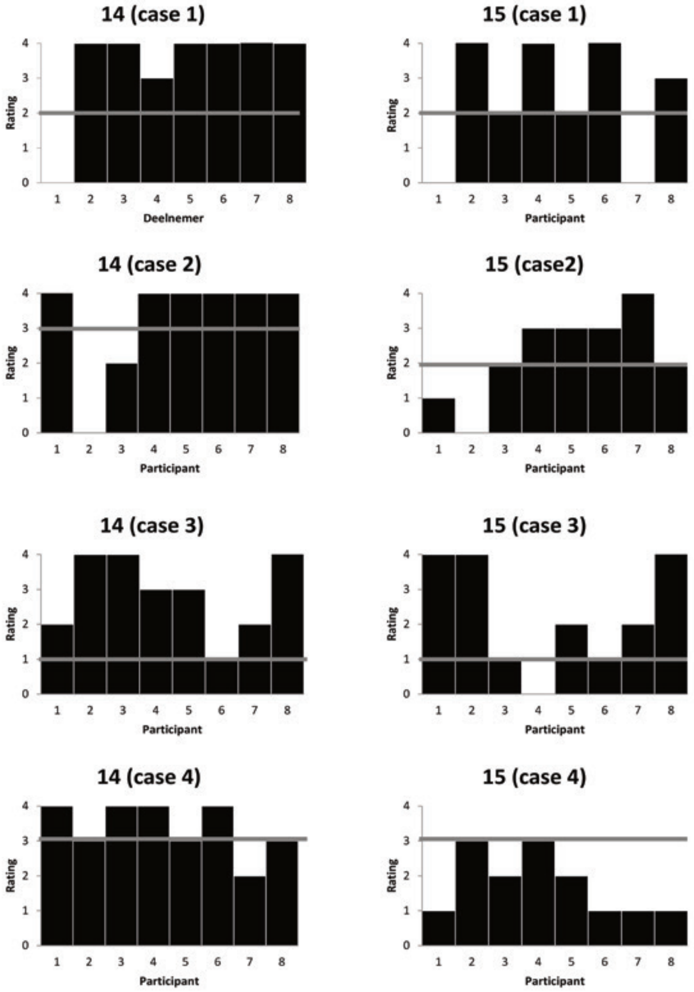


Figure 2 | Assessments of the final assessment questions 14 and 15 per participant per case, expert rating represented by the grey bars

Evaluation of the FIAT-Health 1.0

The topics mentioned by the participants in the evaluation of the FIAT-Health 1.0 are provided in table 2. Most frequently, participants from all user groups found the FIAT-Health contributed to their awareness of the main characteristics of the figure due to its structured approach (n=14). This was particularly frequently mentioned by policy makers (n=5). *Policy maker: "In itself it is useful to systematically assess a figure. It does take a lot of time to assess a figure. It forces one to look at the primary publication again."*

Furthermore, the complexity of the FIAT-Health 1.0 was frequently commented on by policy makers, communication officers and researchers (n=11). *Researcher: "I think it is an interesting tool, because it makes you stop and think about the questions you should ask yourself when reading such a report. But I don't think it is very user friendly, as an Excel file."* The Excel format of the FIAT-Health 1.0 was evaluated as "*time-consuming" (n=9)*. Although two students, a policy maker and a researcher thought the FIAT-Health 1.0 was user-friendly (n=4). The language use was considered complicated (n=7), and some participants (two researchers and one student) could not grasp the goal of the FIAT-Health (n=3). Another topic mentioned in the evaluation was the time investment of checking the primary publications (n=3), while others considered the reference to the primary publication as positive (n=4). Some participants thought the explanations to the questions (in the Dutch version of the FIAT-Health 1.0) were helpful (n=3).

Participants recommended the transformation to an online checklist. Furthermore, some participants commented that not all questions were relevant for the figure they assessed (n=2), or that more in-depth questions regarding for example the methods could be added (n=1). For one participant it was unclear what we meant by 'primary publication'.

Table 2 | Topics in the evaluation of the FIAT-Health 1.0, number of times mentioned.

Topic	Policy Makers	Researchers	Communication officers	Students	Total
High time investment	4	2	2	1	9
The questionnaire/Excel sheet is complex	5	2	3	1	11
Time investment of checking the primary publication	1	2			3
Reference to the primary publication is helpful	2	1	1		4
The structured way of assessing is good for creating awareness of the characteristics of the figure	5	3	3	3	14
Language is complex	1	3	1	2	7
Questions could be more in-depth	1				1
Goal of the questionnaire is unclear		2		1	3
Explanations clarify the question		1		2	3
Not all questions are relevant for the figure		1	1		2
Unclear meaning primary publication				1	1
The FIAT-Health 1.0 is user friendly	1	1		2	4

Consistency of the answers

Out of twenty-six dichotomous questions, seventeen questions were answered consistently among participants. Nine questions we answered inconsistently.

For the following nine questions two or more participants answered inconsistent with the majority of answers:

- *3a, Is the figure expressed in absolute terms?*
- *3c, Does the figure you are assessing match the figure in the primary publication?*
- *4b, Does the definition of the subject of the figure you are assessing match the definition of the subject in the primary publication?*
- *5b, Does the definition of the population of the figure you are assessing match the definition in the primary publication?*
- *7a Is the time period in which the units are counted described in the primary publication?*
- *7b, Does the time period to which the figure applies match the time period in the primary publication?*
- *8a, Are the data on which the figure is based collected periodically?*
- *10a, Were the data collected through an existing registration?* and
- 13a, *Was the figure constructed through modelling?*

Analysis of the numerical assessment questions showed a pattern of inconsistency in how participants responded. On these questions, more than three participants deviated from the majority. Agreement between participant answers per question per case for the dichotomous questions is presented in the supplementary material 1, table S1.1.

FIAT-Health 2.0

Based on the results of the evaluation the FIAT-Health 1.0 was adapted. The questions that were answered inconsistently or unclear by the participants were reformulated and the explanations to specific concepts were specified. Most questions that were answered inconsistently were changed into an open-ended question format, while a few questions on the agreement between the primary publication and the reported figure were revised. In addition, the explanation of one question (nr. 13) was extended.

The construct of the FIAT-Health 1.0, namely the overall quantitative assessment of the figure, was replaced by an open-ended answer format. The new construct of the FIAT-Health is aimed at the systematic answering of questions that are important for the interpretation of a figure on health and healthcare and is no longer aimed at constructing an objective quantitative assessment.

Draft versions of the new FIAT-Health 2.0 were tested by scientific staff at the RIVM and reviewed by the sounding board. Based on their feedback, final adaptions to the language were made, and the last question (15) was changed to assess the 'interpretation of the figure' in the FIAT-Health 2.0, rather than the 'appropriateness of the report of the figure' in the FIAT-Health 1.0. The FIAT-Health 2.0 is presented in table 3. To improve the usability of the instrument a website *www.fiathealth.info* (23) was created. On this website, the instrument can be used with a user-friendly interface, with additional functionalities such as the automatic creation of a summary overview of the main characteristics of a figure based on the responses to the questions.

The FIAT-Health 2.0 consists of factual questions, questions regarding the agreement between the primary publication and the public report, and open-ended assessment questions. The final assessment of the FIAT-Health 2.0 concerns a description of the correctness of the figure and the interpretation of the public report.

Table 3 | The FIAT-Health 2.0

	FIAT-Health 2.0
	What figure would you like to assess? (Provide the phrase in which the figure is mentioned.)
	Question 1. Origin of the figure
1a.	Is the publication in which the figure is reported a primary publication?
1b.	Is the primary publication known?
1c.	Is the primary publication verifiable?
	If the primary publication is not accessible, the FIAT-Health cannot be used. Access to the primary publication is necessary to assess the reported figure with the FIAT-Health 2.0.
1d.	Does the figure you are assessing match the figure in the primary publication?
	Question 2. Credibility of the figure
2a.	Under the responsibility of what institute has the primary publications been published?
2b.	How credible do you consider the author of the primary publication [in relation to this particular figure]?
	Question 3. Expression of the figure
3.	How is the reported figure expressed? (*For example: in absolute terms, percentage, average, fraction*)
	Question 4. Subject to which the figure applies
4a.	To what subject does the reported figure relate?
4b.	Is the subject of the reported figure identical to the subject described in the primary publication?
	Question 5. Population to which the figure applies
5a.	To what population does the reported figure relate?
5b.	Is the population of the reported figure identical to the population described in the primary publication?

Table 3 | Continued

	FIAT-Health 2.0
	Question 6. Geographical area to which the figure applies
6a.	To what geographical area does the reported figure relate?
6b.	Is the geographical area of the reported figure identical to the geographical area described in the primary publication?
	Question 7. Time period to which the figure applies
7a.	To what time period does the reported figure relate?
7b.	Is the time period of the reported figure identical to the time period described in the primary publication?
	Question 8 to 13: Methods of counting and measuring
	Multiple methods of counting and measuring are possible.
	Question 8. Data collection
8.	Are the data on which the figure is based collected only once, periodically or continuously?
	Question 9. Sample
9a.	Is the figure based on a sample?
9b.	What is the sample size?
9c.	What is the response rate?
9d.	Were important groups disregarded in the calculation of the figure? If yes, which groups?
9e.	How do you assess the representativeness of the sample?
	Question 10. Registration
10a.	Were the data collected through an existing registration?
10b.	What registration was used?
10c.	How do you assess the usability of this registration for the calculation of this specific figure?
	Question 11. Survey research
11a.	Were the data collected through survey research?
11b.	Are the questions on which the figure is based described precisely?
11c.	Are the answer categories of the questions described?
11d.	How do you assess the conclusion which was made based on the questions and the answer categories?
	Question 12. Direct observations
12a.	Are the data collected through direct observations?
12b.	How the direct observations take place?
12c.	How do you assess the accuracy of the direct observations?
	Question 13. Modelling
13a.	Was the figure constructed through modelling?
13b.	Are the assumptions which were made in the model known?
13c.	How do you assess the plausibility of the assumptions made in the model?
	Final assessment
	How do you assess the correctness of the figure in the primary publication?
	How do you assess the interpretation of reported figure?

DISCUSSION

The aim of this study was to test and evaluate the FIAT-Health 1.0 amongst its intended user groups, and to compare the assessment of figures with the FIAT-Health to an assessment made by experts.

Qualitative results indicate that the FIAT-Health supports its users to make similar considerations to experts when they assess a publicly reported figure. The participants of this study underlined the value of the structured approach of the FIAT-Health in assessing a figure and noted that it made them consider the figure more critically. Furthermore, the FIAT-Health is considered time-intensive and complex by the participants of this study. The results of this study indicate that it is feasible for potential users to answer factual questions about a figure consistently. Nevertheless, the answers on the quantitative assessment questions were inconsistent.

In line with these results, inconsistently answered and unclear questions of the FIAT-Health 1.0 were rephrased while the consistently answered questions were retained. Most importantly, we revised the underlying construct, in which we assumed that the FIAT-Health can support users in making a quantitative assessment of a figure.

Limitations

The FIAT-Health 1.0 was tested by its intended users. Because of the time-investment participants could only assess one case. As our sample size was small and users did not repeat any measurements, estimates of reliability such as Kappa's (24) or ideally, Krippendorff's Alpha (25) could not be calculated.

As we developed the FIAT-Health 1.0, we might have interpreted the results of its evaluation more positively. By reporting our findings, involving participants outside the researching institute, our preparedness to thoroughly adapt the instrument, and discussing our results with a sounding board outside the project group, we tried to avoid this bias. Furthermore, a risk of selection bias exists due to our purposeful sampling strategy. Those with no interest in using the tool might not have been interested in participating in this study. Seven students declined participation of this study which could indicate that the students might have limited interest in using this tool unless they have a curiosity in healthcare research. Unlike students, policy advisors, communication officers and researchers showed a greater willingness to participate. Consequently, their interest in using a tool to support reporting of figures may be higher.

The evaluation questions were aimed at improving the FIAT-Health, thus participants focussed on what they thought was unclear and could be amended. The positive sides of the FIAT-Health 1.0 might have been underrepresented in their answers.

Context

Most reporting checklists demonstrate a low measure of reliability. Mokkink et al (2010) found a low inter-rater reliability of the quantitative assessment of the COSMIN Checklist (COnsensus-based Standards for the selection of health status Measurement Instruments) (26). In addition, Pieper (2017) who performed a review of systematic reviews using the AMSTAR statement showed low inter-rater reliability as well (27). They concluded that an assessment of instruments using only two reviewers would be insufficient in determining reliability, as raters would use their own subjective judgement. Furthermore, dichotomous items are more likely to be answered reliably than scaled questions (28). It seems to be difficult to construct an objective quantitative assessment of a publication whether it is in science or public communication. Therefore, we consider that in the assessment made using the FIAT-Health, there will always be a certain degree of subjectivity.

While the ratings seemed to be inconsistent, the motivations for assessments of the participants were closely aligned with the motivations provided by the experts. These results support that the FIAT-Health 1.0 did grasp the right items that support the interpretation of a figure. As policy makers and other users indicated that a structured assessment helped them become more aware of the characteristics of the figure, the primary goal of the FIAT-Health, namely supporting interpretation, was reinforced. When we revised the tool, we aimed to further emphasize this goal. To support users in the assessment of figures on health and healthcare, FIAT-Health 2.0 was revised into a qualitative online assessment tool consisting of open-ended questions aimed at a better interpretation of publicly reported figures. Both the FIAT-Health 1.0 scoring instrument and 2.0 assessment tool consist of three types of questions and a final assessment. Questions in the FIAT-Health 1.0 have a closed-ended format, including numerical ratings, while the questions in the FIAT-Health 2.0 primarily have an open-ended format, providing room for descriptive answers and assessments. Both the FIAT-Health 1.0 and 2.0 can be used as a checklist. However, use of the FIAT-Health 2.0 as a checklist is made easier due to its simplified format. The differences between the FIAT-Health 1.0 and 2.0 are described in table 4.

Table 4 | Differences between the FIAT-Health 1.0 and 2.0

	FIAT-Health 1.0	FIAT-Health 2.0
Factual questions	Closed-ended questions on the characterization of the figure answered with 'yes' or 'no'.	Includes both open-ended questions on the characterization of the figure, answered by taking information from the public report, and closed-ended questions, answered with 'yes' or 'no'.
Assessment questions	The user gives a rating on a scale from 1 (negative) to 4 (positive) on methodological aspects.	The user describes his or her assessment on methodological aspects providing a numerical rating.
Questions on the primary publication	Questions on the consistency between the publicly reported figure and the figure as described in the primary publication.	Questions on the consistency between the publicly reported figure and the figure as described in the primary publication are rephrased.
Final assessment	The user rates the 'correctness of the original figure in the primary publication 'and the 'appropriateness of the figure in the report' on a scale from 1 (negative) to 5 (positive).	The user describes his or her considerations on their assessment of the 'correctness of the figure in the primary publication', and the 'interpretation of the reported figure'.
Checklist	Can be used as a checklist.	Use of the FIAT-Health 2.0 as a checklist is made easier due to its simplified format.

Although there are many available checklists and methods to support reporting and assessment of the quality of peer-reviewed scientific publications (14), these checklists that assess statistics in societal publications have not been tested and constructed scientifically. Studies on the use of checklists in peer-reviewed scientific publications indicate that such a checklist does improve the quality of reporting (29). For a long time, lay checklists have been published in the form of popular literature, such as Darrel Huffs book "How to Lie with Statistics" (30). The content of the FIAT-Health 2.0 was constructed systematically. Moreover, the FIAT-Health 2.0 was developed, improved and tested through the involvement of its potential users.

The FIAT-Health 2.0 can contribute to public understanding of statistics in two ways. One, the tool may be used by any person to assess a figure reported in the media. A limitation of this function lies in the construction of the FIAT-Health. We did not have the opportunity to involve the general public in the construction and improvement of the tool, and considering the feedback on the FIAT-Health 1.0, its language might still be difficult to grasp by some. Nevertheless, the tool is publicly available in Dutch and easily accessible online, to be used by those who are interested. Two, the tool is considered useful by policy makers, communication experts and researchers.

These are the people that bring statistics under the attention of the public. If they apply the tool to improve their reporting, we may intervene in the communication flows from those creating the figure (research institutes/ scientific research) to the receivers (the public) (14). The figures may be reported more responsibly including a necessary description of sources, construction and methodology. Improved reporting on the most relevant background characteristics of a figure will give the public the information necessary to interpret the reported figure.

Implications

The potential users of the FIAT-Health have mentioned the usefulness of the tool, indicating that the FIAT-Health would be valuable to the work of policy makers, researchers, and communication officers. Currently, publicly reported statistics are not assessed systematically, but reviewed based on the user's knowledge and expertise. The FIAT-Health 2.0 can help those without expert knowledge to assess statistics systematically or help researchers and communication officers report findings responsibly. Carefully interpreting statistics is time consuming, thus we recommend development of implementation strategies for those who regularly publish statistics. In its current form, the FIAT-Health 2.0 can be used to create a structured overview of the most important characteristics of a figure, or, when short in time, as a simple checklist. Since using a checklist repeatedly is likely to result in better assessments (28), we recommend people to use the FIAT-Health 2.0 frequently.

CONCLUSION

The elements of the FIAT-Health 1.0 were considered useful by the participating policy makers, communication officers and researchers. By transforming the FIAT-Health from a quantitative scoring instrument into an online qualitative assessment tool, the usability of the tool has increased, making it less complex and more suitable for its purpose. The FIAT-Health 2.0 is a unique instrument that has the potential to help policy makers, communication officers and researchers to systematically assess figures, form a structured interpretation of figures, and aid the better reporting of figures on health and healthcare towards the public.

SUPPLEMENTARY MATERIAL 1.

Table S1.1 |Agreement per question per case expressed as number of same answers as part of the total number of given answers.

Nr.	Question	Case 1	Case 2	Case 3	Case 4
1a.	Is the publication in which the figure is reported a primary publication?	7/8	8/8	8/8	8/8
1b.	Is the primary publication known?	7/7	7/8	7/8	5/7*
1c.	Is the primary publication verifiable?	7/7	8/8	7/8	4/7*
3a.	Is the figure expressed in absolute terms?	4/8*	6/8*	5/8*	6/8*
3b.	Is the figure expressed in relative terms?	6/8*	6/6	5/5	5/6
3c.	Does the figure you are assessing match the figure in the primary publication?	5/7*	6/8*	5/8*	8/8
4b.	Does the definition of the subject of the figure you are assessing match the definition of the subject in the primary publication?	5/7*	6/8*	5/8*	5/8*
5b.	Does the definition of the population of the figure you are assessing match the definition in the primary publication?	6/7	5/8*	6/8*	5/8*
6b.	Does the geographical area of the figure you are assessing match the geographical area in the primary publication?	6/7	5/8*	8/8	8/8
7a	Is the time period in which the units are counted described in the primary publication?	8/8	6/8*	7/8	5/8*
7b.	Does the time period to which the figure applies match the time period in the primary publication?	6/7	4/8*	7/8	6/8*
8a.	Are the data on which the figure is based collected periodically?	8/8	5/8*	6/8*	7/8
8b.	Are the data on which the figure is based collected only once?	N.A.	5/5	2/2	1/1
9a.	Is the figure based on a sample?	8/8	5/8*	7/8	7/8
9b.	Is the sample size known?	7/8	3/3	1/1	1/1
9c.	Is the response known?	8/8	2/3	1/1	1/1
9d.	Were important groups disregarded in the calculation of the figure?	7/8	2/3	1/1	N.A.
10a.	Were the data collected through an existing registration?	7/8	7/8	5/8*	5/8*
10b.	Is it known which registration was used?	1/1	7/7	2/3	3/3
11a.	Were the data collected through survey research?	8/8	6/8*	7/8	7/8
11b.	Are the questions on which the figure is based described precisely?	7/8	5/6	1/1	1/1
11c.	Are the answer categories of the questions described?	7/8	6/6	1/1	1/1
12a.	Are the data collected through direct observations?	8/8	8/8	8/8	6/8*
12b.	Is it known how the direct observations took place?	N.A.	N.A.	N.A.	2/2
13a.	Was the figure constructed through modelling?	7/8	5/8*	8/8	6/8*
13b.	Are the assumptions which were made in the model known?	1/1	2/3	8/8	3/6*

** answers are inconsistent*
N.A. not applicable as no answers were given

REFERENCES

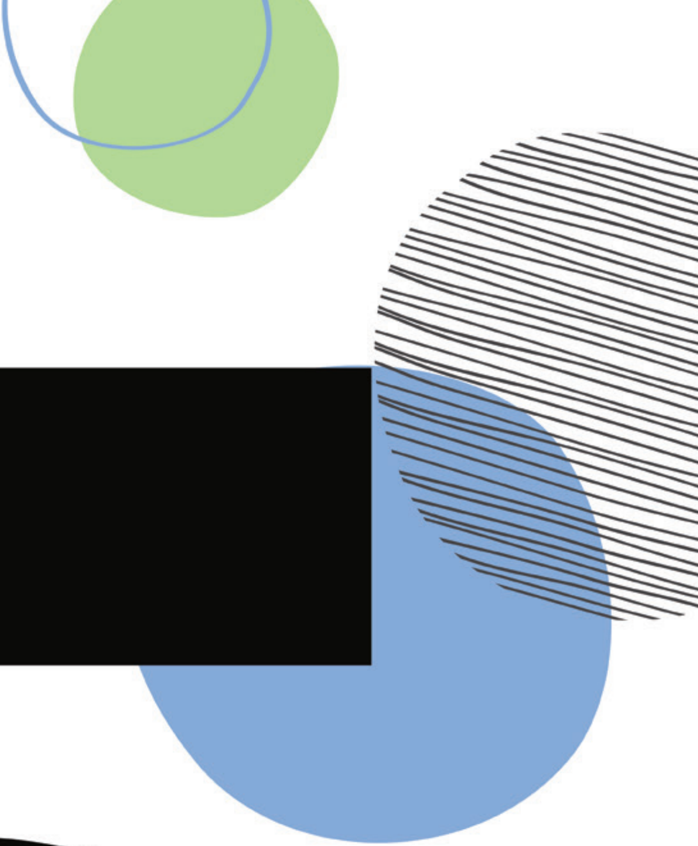
1. Young ME, Norman GR, Humphreys KR. Medicine in the Popular Press: The Influence of the Media on Perceptions of Disease. Plos One. 2008;3(10).
2. Weingart P. Science and the media. Res Policy. 1998;27(8):869-79.
3. Zebregs S, van den Putte B, Neijens P, de Graaf A. The Differential Impact of Statistical and Narrative Evidence on Beliefs, Attitude, and Intention: A Meta-Analysis. Health Communication. 2015;30(3):282-9.
4. Niederdeppe J, Roh S, Dreisbach C. How Narrative Focus and a Statistical Map Shape Health Policy Support Among State Legislators. Health Commun. 2016;31(2):242-55.
5. Moreland-Russell S, Harris JK, Israel K, Schell S, Mohr A. "Anti-Smoking Data Are Exaggerated" Versus "The Data Are Clear and Indisputable": Examining Letters to the Editor About Tobacco. Journal of Health Communication. 2012;17(4):443-59.
6. Frost K, Frank E, Maibach E. Relative risk in the news media: A quantification of misrepresentation. Am J Public Health. 1997;87(5):842-5.
7. Yavchitz A, Boutron I, Bafeta A, Marroun I, Charles P, Mantz J, et al. Misrepresentation of Randomized Controlled Trials in Press Releases and News Coverage: A Cohort Study. Plos Medicine. 2012;9(9).
8. Black N. Evidence based policy: proceed with care. BMJ. 2001;323(7307):275-9.
9. Simmerling A, Janich N. Rhetorical functions of a 'language of uncertainty' in the mass media. Public Underst Sci. 2015.
10. Caulfield T. The Commercialisation of Medical and Scientific Reporting. PLoS Med. 2005;1(3):e38.
11. Sato H. Agenda Setting for Smoking Control in Japan, 1945-1990: Influence of the Mass Media on National Health Policy Making. Journal of Health Communication. 2003;8(1):23-40.
12. Furedi A. The public health implications of the 1995 'pill scare'. Hum Reprod Update. 1999;5(6):621-6.
13. Spiegelhalter D. Trust in numbers. Journal of the Royal Statistical Society Series A: Statistics in Society. 2017;180(4):948-65.
14. Altman DG, Simera I, Hoey J, Moher D, Schulz K. EQUATOR: reporting guidelines for health research. Lancet. 2008;371.
15. Stevens GA, Alkema L, Black RE, Boerma JT, Collins GS, Ezzati M, et al. Guidelines for Accurate and Transparent Health Estimates Reporting: the GATHER statement. PLoS Med. 2016;13(6):e1002056.
16. Walker N, Bryce J, Black RE. Interpreting health statistics for policymaking: the story behind the headlines. The Lancet. 2007;369(9565):956-63.
17. Dobbins M, Jack S, Thomas H, Kothari A. Public health decision-makers' informational needs and preferences for receiving research evidence. Worldviews Evid Based Nurs. 2007;4(3):156-63.
18. Oliver K, Innvar S, Lorenc T, Woodman J, Thomas J. A systematic review of barriers to and facilitators of the use of evidence by policymakers. BMC Health Serv Res. 2014;14:2.
19. von Roten FC. Do we need a public understanding of statistics? Public Understanding of Science. 2006;15(2):243-9.
20. Gerrits RG, Kringos DS, van den Berg MJ, Klazinga NS. Improving interpretation of publically reported statistics on health and healthcare: the Figure Interpretation Assessment Tool (FIAT-Health). Health Res Policy Syst. 2018;16(1):20.

21. Hampshaw S, Cooke J, Mott L. What is a research derived actionable tool, and what factors should be considered in their development? A Delphi study. BMC Health Services Research. 2018;18(1):740.
22. Feinstein AR. Clinimetrics: Yale University Press; 1987.
23. FIAT-Health 2.0. https://www.fiathealth.info/. Accessed 22 October 2018. [
24. McHugh ML. Interrater reliability: the kappa statistic. Biochemia Medica. 2012;22(3):276-82.
25. Hayes KF, Krippendorff K. Answering the Call for a Standard Reliability Measure for Coding Data. COMMUNICATION METHODS AND MEASURES,. 2007;1(1):77-89.
26. Mokkink LB, Terwee CB, Gibbons E, Stratford PW, Alonso J, Patrick DL, et al. Inter-rater agreement and reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) checklist. BMC Med Res Methodol. 2010;10:82.
27. Pieper D, Jacobs A, Weikert B, Fishta A, Wegewitz U. Inter-rater reliability of AMSTAR is dependent on the pair of reviewers. BMC Medical Research Methodology. 2017;17:98.
28. Oremus M, Oremus C, Hall GBC, McKinnon MC. Inter-rater and test–retest reliability of quality assessments by novice student raters using the Jadad and Newcastle–Ottawa Scales. BMJ Open. 2012;2(4).
29. Han S, Olonisakin TF, Pribis JP, Zupetic J, Yoon JH, Holleran KM, et al. A checklist is associated with increased quality of reporting preclinical biomedical research: A systematic review. PLoS ONE. 2017;12(9):e0183591.
30. Huff D. How to Lie With Statistics. New York: W. W. Norton & Company; 1954.

PART II

RESPONSIBLE REPORTING OF HEALTH SERVICES RESEARCH

5



The occurrence and nature of questionable research practices in the reporting of messages and conclusions in international scientific Health Services Research publications: A structured assessment of publications authored by researchers in the Netherlands

AUTHORS:

Reinie G. Gerrits, Tessa Jansen, Joko Mulyanto, Michael J. van den Berg, Niek S. Klazinga, Dionne S. Kringos

Published in BMJ Open 2018;9:e027903

ABSTRACT

OBJECTIVES Explore the occurrence and nature of questionable research practices (QRPs) in the reporting of messages and conclusions in international scientific HSR publications authored by researchers from health services research (HSR) institutions in the Netherlands.

DESIGN In a joint effort to assure the overall quality of HSR publications in the Netherlands, thirteen HSR institutions in the Netherlands participated in this study. Together with these institutions, we constructed and validated an assessment instrument covering 35 possible QRPs in the reporting of messages and conclusions. A QRP in the reporting of messages and conclusions in HSR is defined as *"to report, either intentionally or unintentionally, conclusions or messages that may lead to incorrect inferences and do not accurately reflect the objectives, the methodology or the results of the study."* Two reviewers independently assessed a random sample of 116 HSR articles authored by researchers from these institutions published in international peer-reviewed scientific journals in 2016.

MAIN OUTCOME MEASURES Median number of QRPs per publication, the percentage of publications with observed QRP frequencies, occurrence of specific QRPs, and difference in total number of QRPs by methodological approach, type of research, and study design.

RESULTS We identified a median of six QRPs per publication, out of 35 possible QRPs. QRPs occurred most frequently in the reporting of implications for practice, recommendations for practice, contradictory evidence, study limitations, and conclusions based on the results and in the context of the literature. We identified no differences in total number of QRPs in papers based on different methodological approach, type of research or study design.

CONCLUSIONS Given the applied nature of HSR, both the severity of the identified QRPs, and the recommendations for policy and practice in HSR publications warrant discussion. We recommend that the HSR field further define and establish its own scientific norms in publication practices to improve scientific reporting and strengthen the impact of HSR. The results of our study can serve as an empirical basis for continuous critical reflection on the reporting of messages and conclusions.

INTRODUCTION

In 2009, Chalmers and Glasziou estimated that 85% of research funding in biomedical sciences was wasted avoidably (1), resulting in *The Lancet*'s series "Increasing value: reducing waste". This series has stirred the international scientific community, prompting funders, regulators, academic institutions, and scientific publishers to act. Funders of biomedical research have responded by organising conferences on research waste, and journal editors have initiated discussions on data sharing and open access (2). While evidence for questionable research practices (QRPs) in biomedical sciences is mounting (1), little is known about the occurrence and nature of QRPs in the policy- and management-oriented field of health services research (HSR). In particular, QRPs in the reporting of messages and conclusions have flown under the radar. The term 'questionable research practices' is commonly used to describe practices such as selective publication of results, concealing of conflicts of interests, and describing a hypothesis after finding significant results (3). A questionable practice is not necessarily wrongful, but does 'raise questions'. In this study we further define the meaning of questionable research practices in the reporting of messages and conclusions in the field of HSR specifically.

The HSR field is an applied field of research, and produces evidence on topics such as co-payments, evaluation of quality improvement efforts, cost-effectiveness of medications, patient empowerment, therapy compliance, and effects of policies. Given the growing evidence for the prevalence of QRPs in the reporting of messages and conclusions in the biomedical field (4, 5), QRPs may also occur in the HSR field. In the biomedical field, a systematic review by Chiu et al. (2017) shows that estimates for the occurrence of questionable research practices in the interpretation of results in scientific publications vary from 10% of publications deriving discordant conclusions from study results to 100% of publications containing rhetorical practices resulting in spin, such as failure to compare risk to benefits in randomized controlled trials (4).

Just like biomedical researchers, health services researchers are under pressure to publish in high-impact journals to increase their citation scores and attract media attention to augment their prestige and chances for future research funding and job security (6-9). Unlike biomedical research, HSR findings are not easily generalised from one local or national health services setting to another, and messages and conclusions tend to be limited to a specific national context (10). A broad spectrum of quantitative and qualitative methods is used in HSR, including designs that are less subject to strict codes of execution than randomized controlled trials, such as observational and case study designs. Furthermore, HSR has difficulty creating alignment between the construction of scientific knowledge and the implementation of that knowledge in policy and practice (11). This combination of HSR specific characteristics may result in a different set of QRPs in the reporting of a scientific study. The variation of designs other than RCTs, as is more common in the biomedical field, might invite unjustified claims of causality. Moreover, the context specific research may increase unjustified claims of generalisability, and the difficulty in translating knowledge to practice may result in unsupported recommendations or implications.

Although reporting in scientific publications is highly standardised, the discussion and conclusion sections offer researchers relative freedom when deriving messages and conclusions from study results (5). We explored the occurrence and nature of QRPs in the reporting of messages and conclusions in international scientific HSR publications authored by researchers from HSR institutions in the Netherlands. We also examined the relationship between study type, methodology, and design and the occurrence of QRPs. With our study, we want to fuel the debate on fostering responsible messages and conclusions, and provide a basis for the discussion on QRPs in the international HSR field.

METHODS

Setting

This study assessed scientific publications authored by researchers from 13 HSR groups, departments, or institutions (hereafter referred to as "HSR institutions") in the Netherlands, including both academic and non-academic institutions. These institutions all agreed to participate in an effort to assure the overall quality of HSR publications in the Netherlands.

Defining QRPs in the reporting of messages and conclusions in HSR

We conducted a literature review on QRPs in the reporting of messages and conclusions in biomedical research and HSR (12-14). An initial definition of QRPs in the reporting of messages and conclusions in HSR was proposed and discussed at a consensus meeting with the directors/leaders of the 13 participating institutions. This was then validated through inputs from five leading international health services researchers (10 were invited; 50% non-response), and resulted in the following amended definition:

"To report, either intentionally or unintentionally, conclusions or messages that may lead to incorrect inferences and do not accurately reflect the objectives, the methodology or the results of the study."

Measurement instrument

We developed an extensive list of QRPs in the reporting of messages and conclusions. Items were based on the EQUATOR checklists (15) and earlier checklists for identifying "spin" (ie, "a way to distort science reporting without actually lying") (5) or other QRPs (13, 14, 16, 17). The proposed list of QRPs was reviewed, refined, and complemented using 14 semi-structured interviews with the directors/leaders and representatives (n=19) of the 13 participating HSR institutions. Next, the five participating international health services researchers provided email feedback on the list resulting from these interviews; the list was adapted accordingly, resulting in 35 possible QRPs in the reporting of messages and conclusions in HSR publications.

We developed a data extraction form in Excel that contained the list of QRPs and bibliometric information, and conducted a pilot to evaluate its feasibility and usability. In the pilot, two assessors (RG, TJ) independently assessed five international HSR publications to identify modifications needed to improve the form, and to align the interpretation of the items. The project group discussed the proposed modifications, resulting in the final version. The data extraction form, (supplementary material 1) and a methodology of the development of the data extraction form (supplementary material 2) is provided in the supplementary material.

Sample

We aimed to include ten HSR publications from each participating HSR institution. Inclusion criteria were: published in 2016 in an international peer-reviewed scientific journal, written in English, reporting HSR findings, and first- and/or last-authored by researchers affiliated with the respective HSR institution. As both the first author and the research institution are likely important factors influencing the occurrence of QRPs, only unique first authors were included in the publication. Moreover, not more than 10 publications per institution were included. This will ensure a maximum spread of authors and institutions across the sample.

Publication lists of the HSR institutions were retrieved either by searching publicly accessible online sources (eg, annual reports, open repositories or the research groups' website) or obtained from secretaries or librarians. All lists were verified by the respective HSR institutions. These lists included both HSR and non-HSR publications.

Two researchers (RG, TJ) selected all titles from the 13 publication lists that were likely to indicate empirical or systematic assessment studies in HSR. Publications were included if their title fitted the definitions of HSR by Juttmann (2007) (18) and Lohr & Steinwachs (2002) (19). These definitions are commonly used by HSR institutions (eg, in education) in the Netherlands. To select HSR studies, TJ and RG first individually selected titles from the publication lists. Next, RG and TJ compared their selections of titles and noted any differences. After completing the selection of the first HSR publications, selection was reviewed and approved by the research group (NK, DK, MB). TJ and RG then continued applying the selection method to the remaining publication lists. In a consensus meeting between TJ and RG, differences in selected titles were resolved by discussing its fit with the definition. Consensus was reached on all included publications.

The HSR publications (n=717) were assigned a random number. Per institution, the publications with unique first authors with the lowest assigned number were included in the sample. Three HSR institutions did not have enough publications with unique first authors, resulting in a selection of nine, eight, and two publications for these institutions. Furthermore, two publications were excluded during assessment because they concerned research protocols. These publications were replaced by another publication authored by the same institution. One publication was excluded because its methodology was considered incomprehensible by the reviewers. Ultimately, 116 HSR publications were included (16% of tot sample).

Assessment process

Two reviewers independently assessed all publications (RG and TJ or RG and JM). RG has primarily qualitative HSR experience and is trained in health economics. TJ and JM have primarily quantitative HSR experience and are trained in public health, management, economics, and law; and medicine, respectively.

The assessment started with a test phase. During this phase, agreements and disagreements in assessments of the first 30 publications were thoroughly discussed (by RG, TJ, NK, and DK) to increase the accuracy of the assessments; agreement between the two reviewers (TJ, RG) was 81% for the first 20 publications, which increased to 82% when assessing the next 10 publications. The notion emerged that it was necessary having two reviewers with complementary expertise assess each publication independently, followed by a consensus procedure and random check by the project leaders. RG trained the third reviewer (JM).

RG assessed all included publications, while TJ assessed the first 59 publications, and JM the remaining 57. All data were entered in the data extraction form. QRPs were coded as either 1, “present”; 0, “not present”; -8, “not applicable to this study” (primarily used for items not applicable for qualitative research); or -9, “not assessable”. To justify their assessments, the reviewers recorded their motivation for every identified QRP. At a later stage, QRPs in implications and recommendations for policy and practice were further refined into “not mentioned” if no implication or recommendation was included in the publications, and “not sufficiently justified”, if the authors did not provide any explanation for their implications or recommendations. The reviewers held regular consensus meetings (after review of 10 publications) to discuss and reach agreement on all identified QRPs.

During the consensus meetings, the reviewers compared their assessment of all items. Inconsistencies between the individually assessed QRPs were identified, discussed and adapted. Any remaining disagreements (n=2) were resolved by a senior researcher (DK). NK and DK each reassessed a random sample of six publications, so 10% of all included publications (n=12). As a result, two identified QRPs were retracted, and two QRPs were added to the reassessed publications.

Analysis

The characteristics of the included publications were described by calculating their occurrence with the percentage or mean number of publications.

We counted the total number of QRPs per publication, and the percentage of HSR publications with number of observed QRPs. The latter was visualised in a histogram. Occurrence of specific QRPs was calculated as a percentage of publications containing this particular QRP. The percentage of publications containing QRPs that were not applicable to qualitative research was calculated only for quantitative and mixed-methods-based publications (n = 83), (e.g. the QRP: “The relevance of statistically significant results with small effect size is overstated” is only applicable to quantitative research).

We used a Kruskal-Wallis test to calculate the difference in total number of QRPs applicable to all research designs by methodological approach (quantitative, qualitative, and mixed), type of research (descriptive, exploratory, hypothesis testing, and measurement instruments), and study design (observational, (quasi) experimental, systematic review, economic evaluation, case study, and meta-analyses). We used the STROBE checklist for observational studies in the reporting of this research (20). Analyses were conducted using SPSS version 24 (21).

Patient and Public Involvement

No patients were involved in this study. This study was designed with the input provided by the participating HSR institutions at a consensus meeting at the onset of the study, and individual interviews with the directors/leaders of the 13 participating institutions. During a progress meeting with the participating institutions, preliminary (aggregated level) results were discussed to validate and complement the interpretation of findings.

Ethics approval

A waiver for ethical approval was obtained for this study from the medical ethics review committee at Amsterdam UMC. To avoid negative consequences for the authors of the included publications, each publication was assigned a unique identification number. Extracted data were entered in SPSS using this number to separate author information from the study data.

RESULTS

Characteristics of included publications

Table 1 presents the characteristics of the 116 included publications from the 13 participating HSR institutions. To summarise, 54.3% of the publications were quantitative, 28.4% were qualitative, and 17.2% applied a mixed-methods approach. Sixteen percent of the publications were based on a published study protocol. The mean impact factor of the journals was 2.81, and the average number of authors was six.

Occurrence of QRPs per publication

Of the 116 HSR publications, the median number of QRPs per publication was six (interquartile range, 5·75), out of 35 possible QRPs. The distribution of the observed frequency of QRPs across publications is visualised in figure 1.

Table 1 | Characteristics of included publications

Total (N= 116)		n (%)
HSR domain	Policy	19 (16·4)
	Social factors	11 (9·5)
	Financing Systems	10 (8·6)
	Organizational structures & processes	43 (37·1)
	Health technologies	11 (9·5)
	Personal Behaviours	22 (19·0)
Methodological approach	Quantitative	63 (54·3)
	Qualitative	33 (28·4)
	Mixed methods	20 (17·2)
Type of research	Descriptive	31 (26·7)
	Exploratory	59 (50·9)
	Hypothesis testing	19 (16·4)
	Measurement instruments	5 (4·3)
	Other	2 (1·7)
Design	Observational	59 (50·9)
	(Quasi) experimental	9 (7·8)
	Systematic review	17 (14·7)
	Economic evaluation	5 (4·3)
	Meta analyses	3 (2·6)
	Case study	22 (19·0)
	Other	1 (0·9)
Protocol published		19 (16·4)
Funder of study stated		98 (84·5)
Contributions stated		57 (49·1)
Number of included journals		80 (100·0)
		Mean
Impact factor journal (n=93 publications*)		2·81 (SD 1·45)
Number of authors (n=116)		6·12 (SD 5·53)

** Not all journals had an impact factor. Mean impact factor was calculated over 93 publications.*

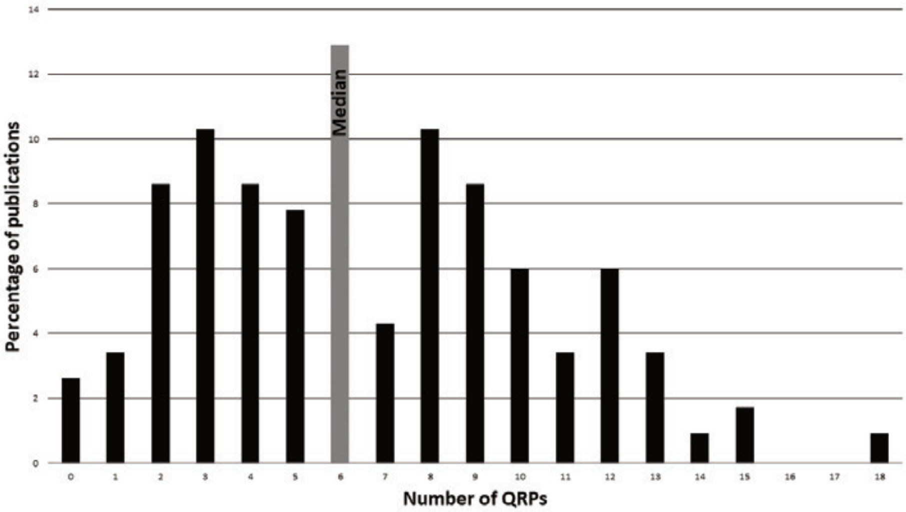


Figure 1 | Percentage of HSR publications with number of observed QRPs in the reporting of messages and conclusions

Frequency of QRPs per type

For each of the QRPs, we counted how often they were identified in the included publications. Supplementary material 3, table S3.1 presents the percentage of occurrence per QRP type.

QRPs that occurred most frequently were:

- Implications for policy and practice do not adequately reflect the results in the context of the referenced literature (69·0%)*;
    *In 50·0% of publications, no implications for policy and practice were mentioned, and in 19·0% of publications, implications were mentioned without adequate justification.
- Recommendations for policy and practice do not adequately reflect the results in the context of the referenced literature (65·5%)**;
    **In 34·5% of publications, no recommendations for policy and practice were reported, and in 31·0% of publications, recommendations were mentioned without adequate justification.
- Contradicting evidence is poorly documented (63·8%);
- Conclusions do not adequately reflect the findings as presented in the results section (46·6%);
- Possible impact of the limitations on the results is not or poorly discussed (44·0%);
- Conclusions are not supported by the results as presented in the context of the referenced literature (43·1%).

QRPs that occurred least frequently were:

- The main source of evidence for supporting the results is based on the same underlying data (2·6%);
- Generalising findings to populations not included in the original sample is not justified (2·6%);
- Causative wording is used in the hypothesis/research question, although there is no theory to support causation (2·4%);
- Possible clinical relevance of statistically non-significant results is not addressed (2·4%);
- Generalising findings to time periods not included in the original study is not justified (0·0%).

Distribution of QRPs

Figure 2 shows the distribution of QRPs across publications. The horizontal axis shows the publications (n=116) ordered from the publication with the lowest (0) to the highest number (18) of observed QRPs in the reporting of messages and conclusions. The vertical axis shows the QRPs ordered from least (Generalisation to different time period) to most (Implications for practice are lacking) frequently observed. On the right vertical axis, the occurrence of QRPs is presented in number of QRPs counted. Each dot represents a QRP.

The difference in the number of QRPs by publication characteristics

Table 2 shows the associations between total number of QRPs (applicable to all study designs) and methodological approach (quantitative, qualitative, and mixed), type of research (descriptive, exploratory, hypothesis testing, and measurement instruments), and study design (observational, (quasi) experimental, systematic review, economic evaluation, case study, and meta-analyses). No statistically significant differences in number of QRPs was found by type of research, methodological approach, or study design.

*The full QRP name is provided in supplementary material 3, table 1 ordered from least frequently found (Generalisation to different time period) to most frequently found (Implications for practice are lacking) QRP.'

Figure 2 | Distribution of QRPs in the reporting of messages and conclusions across HSR publications, ordered from lowest to highest number of observed QRPs.

Table 2 | Association between total number of QRPs and type of research, methodological approach, and study design

	Median	95% CI	p-value
Methodological approach			0·339
Quantitative	5	4·88 – 6·43	
Qualitative	6	4·98– 7·62	
Mixed methods	7	5·34 – 8·46	
Type of research			0·295
Descriptive	6	4·77 – 6·78	
Exploratory	7	5·76 – 7·60	
Hypothesis testing	4	3·40 – 6·81	
Measurement instruments	5	2·14 – 6·66	
Other	5	-33·12 – 43·12	
Study design			0·159
Observational	6	5·56 – 7·21	
(Quasi) experimental	3	2·07 – 5·71	
Systematic review	6	4·61 – 8·33	
Economic evaluation	4	1·61 – 7·59	
Case studies	6	4·71 – 8·01	
Meta-analyses	5	0·50 – 10·84	

DISCUSSION

We explored the occurrence and nature of QRPs in the reporting of messages and conclusions in international scientific HSR publications authored by researchers from HSR institutions in the Netherlands, and examined the relationship between study type, methodology, and design and the occurrence of QRPs. Our results indicate that HSR publications have a median of six QRPs per publication. We identified most QRPs in the reporting of implications for policy and practice, recommendations for policy and practice, contradictory evidence, study limitations, and conclusions as based on the results and in the context of the literature. No significant associations between number of QRPs and type of study, study design, or methodological approach were identified.

Limitations and Strengths

We applied a broad and sensitive definition of 'questionable', for instance by considering the absence of contradictory evidence or the absence of implications and recommendations for policy and practice as a QRP. The choice to not present contradictory evidence does not defy current publication checklists, yet this practice may hinder interpretation of findings in the full context of evidence. If authors searched for contradictory evidence, but did not mention its absence, readers of the publication would not have any clues on its existence.

Knowledge on the occurrence of QRPs is often derived from survey studies, relying on self-report (3). These studies focus on the knowledge of consciously conducted, well-known QRPs. Our assessment approach allowed us to gain insight in less severe, more likely unconsciously occurring QRPs in the reporting of messages and conclusions specifically. The number of QRPs identified through assessment is generally higher than in studies relying on self-report (3, 4). With our broad definition encompassing 35 possible QRPs we bring to light the areas that offer possibilities for further enhancing publication practices in HSR. Consequently, this definition allows for a discussion in the field of HSR on the extent to which the identified QRPs are acceptable. This is an important strength of our applied approach.

Although we endeavoured to develop a reliable measurement instrument that would guide the review process, the instrument allowed latitude for the reviewer's interpretation. Consequently, a different group of reviewers might arrive at somewhat different scoring frequencies for observed QRPs. However, because we defined each QRP in detail, it is unlikely that there would be substantial differences in the overall distribution of different types of QRPs across publications. Our consensus method contains a degree of subjectivity, and there is the risk that one reviewer's opinion will dominate. To counteract this, NK and DK performed random checks on 10% of all assessments. By recording the motivation for every identified QRP, we supported the consistency of our measurement and justified our results. Because publications were selected based on the title, selection bias might have occurred. Considering we found no relationship between study characteristics and number of QRPs, it is unlikely that a different sample would have led to different results. Inevitably, reviewers sometimes assessed publications written by authors they knew professionally or personally, and as such, a positive view of a colleague's work might have led to underestimating the QRPs in these publications.

Our study results may be representative for HSR research publications internationally. Given the fact that publication in international journals is highly standardised in terms of language (English) and format, our findings can most likely be transferred to HSR communities in other countries.

Interpretation

In HSR publications, recommendations for policy and practice warrant most attention. A study can be conducted properly, using a sound design and appropriate methodology. However, making recommendations without adequate justification could lead to incorrect inferences in policy and the management of healthcare, and undermine society's confidence in science (11, 22-25).

Measures for safeguarding scientific soundness like those often used in biomedical research (eg, trial registration, open data policies, and an improved reporting and archiving infrastructure (26)) do not address reporting conclusions not supported by study results, and are not tailored to the observational and explorative designs most prevalent in HSR. Moreover, existing publication checklists address a report's completeness, but do not question the justification of the conclusions (5). If we intend to improve the reporting of HSR conclusions and recommendations, we will need to better understand the factors that influence authors when reporting the discussion and conclusions section of an HSR publication eg, media pressure and relationships with funders (6-8, 27). Journals may have influence on the reporting of a study through control of the review process (28). Moreover, research institutions may prevent the occurrence of QRPs by enhancing internal integrity, training in scientific writing and communication amongst researchers (29). Consequently, subsequent research can focus on what influences researchers when writing their scientific publications, and what factors play a role in the process from research design to the acceptance of a manuscript by a peer-reviewed journal.

A third of the HSR publications studied gave no recommendations for policy or practice, while another third did not provide an adequate justification for the recommendations. One could argue that HSR is an applied field of research, and that its ultimate goal should be to contribute to better health services and systems; researchers should therefore take responsibility for providing guidance to those who can act on the research findings instead of leaving them empty-handed. On the other hand, health services researchers may feel more comfortable committing to a more traditional interpretation of the role of academics, refraining from normative judgement. If the latter is the dominant viewpoint, the HSR community needs to consider the role of scientific evidence in helping decision-makers address the challenges they face, and informing policies and practices. Internationally, the HSR community has been promoting further strengthening of the link between HSR and practice (30).

In biomedical research, research being "new" might contribute to a confused assessment of implications (31). This problem is amplified in HSR, where there is a limited accumulation of evidence. HSR considers a larger range of contextual factors and stakeholders in politics or management. Moreover, HSR recommendations are often based on observational or exploratory research, which is considered to be weak evidence in biomedical circles (eg, the GRADE checklist) (32). Perhaps the norms determined by the biomedical research field make health services researchers hesitant to provide any implications or recommendations at all.

Implications and recommendations for policy and practice

The HSR field currently seems to adhere to the norms and expectations set by the biomedical field, even though HSR is multidisciplinary, and differences in approach and type of methodology pose serious challenges to observing these norms. Therefore, the HSR community needs to further define specific scientific norms appropriate to the field.

Scientific norms are developed through the forum of a scientific community (33). This forum function is particularly strong in the Netherlands, where a community of HSR institutions work together closely. Our study was able to bring together the main Dutch academic and non-academic HSR institutions. Consequently, the results of our study help to facilitate critical reflection on the current state of research and encourage debate on how to systematically advance the reporting of messages and conclusions in HSR. Such a debate in the Dutch context is needed, given the attempts over the past decade by the Netherlands Organisation for Health Research and Development (ZonMw) to strengthen the link between research and practice. It would also be very timely, considering the ongoing, overarching Dutch research programme on responsible research practices funded by ZonMw, of which this study is a part. We recommend the HSR community to reflect on the questions our results bring forward: how do we include implications and recommendations for policy and practice in scientific publications?; how should we describe conclusions in context of literature with limited accumulation of evidence?; and what is the severity of the identified QRPs? Through this publication, we would like to urge journal editors and those working in the international field of HSR to join in this debate. After establishing norms regarding these frequently occurring QRPs, journal editors and HSR institutions may contribute to the prevention of QRPs by implementing strategies tailored to HSR research specifically.

CONCLUSIONS

QRPs in the reporting of messages and conclusions occur frequently in peer-reviewed international scientific HSR publications from Dutch HSR institutions. These QRPs differ in severity and cannot always be qualified as wrongful, but they do 'raise questions'. To ensure the applicability of HSR research in policy and practice, the HSR field should reflect on scientific norms for the reporting of conclusions and the inclusion of recommendations for policy and practice. Our study can serve as an empirical basis for continuous critical reflection on the current state of research, and encourage debate on how to systematically advance the reporting of messages and conclusions in HSR.

REFERENCES

1. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. The Lancet. 2009;374(9683):86-9.
2. Moher D, Glasziou P, Chalmers I, Nasser M, Bossuyt PMM, Korevaar DA, et al. Increasing value and reducing waste in biomedical research: who's listening? The Lancet. 2016;387(10027):1573-86.
3. Fanelli D. How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. PLOS ONE. 2009;4(5):e5738.
4. Chiu K, Grundy Q, Bero L. 'Spin' in published biomedical literature: A methodological systematic review. PLoS Biol. 2017;15(9):e2002173.
5. Boutron I, Ravaud P. Misrepresentation and distortion of research in biomedical literature. Proc Natl Acad Sci U S A. 2018;115(11):2613-9.
6. Ioannidis JPA, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. The Lancet. 2014;383(9912):166-75.
7. Weingart P. Science and the media. Res Policy. 1998;27(8):869-79.
8. Bouter LM. Commentary: Perverse Incentives or Rotten Apples? Accountability in Research. 2015;22(3):148-61.
9. Tijdink JK, Schipper K, Bouter LM, Maclaine Pont P, de Jonge J, Smulders YM. How do scientists perceive the current publication culture? A qualitative focus group interview study among Dutch biomedical researchers. BMJ Open. 2016;6(2):e008681.
10. Porter ME. What Is Value in Health Care? New England Journal of Medicine. 2010;363(26):2477-81.
11. Groen TP. Startdocument 'Systeemfalen' The Hague; 2013.
12. Horton R. The rhetoric of research. BMJ. 1995;310(6985):985.
13. Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. JAMA. 2010;303(20):2058-64.
14. Ochodo EA, de Haan MC, Reitsma JB, Hooft L, Bossuyt PM, Leeflang MM. Overinterpretation and misreporting of diagnostic accuracy studies: evidence of "spin". Radiology. 2013;267(2):581-8.
15. Altman DG, Simera I, Hoey J, Moher D, Schulz K. EQUATOR: reporting guidelines for health research. Lancet. 2008;371.
16. Stevens GA, Alkema L, Black RE, Boerma JT, Collins GS, Ezzati M, et al. Guidelines for Accurate and Transparent Health Estimates Reporting: the GATHER statement. PLoS Med. 2016;13(6):e1002056.
17. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care. 2007;19(6):349-57.
18. Juttmann RE. Het domein van gezondheidszorgonderzoek. In: Plochg T, Juttmann RE, Klazinga NS, Mackenbach JP, editors. Handboek gezondheidszorgonderzoek. Houten: Bohn Stafleu van Loghum; 2007. p. 22-3.
19. Lohr KN, Steinwachs DM. Health Services Research: An Evolving Definition of the Field. Health Serv Res. 2002;37(1):15-7.
20. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, et al. [The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies]. Rev Esp Salud Publica. 2008;82(3):251-9.

21. IBM Corp. Released 2016. IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp.
22. Lavis JN, Ross SE, Hurley JE, Hohenadel JM, Stoddart GL, Woodward CA, et al. Examining the role of health services research in public policymaking. Milbank Q. 2002;80(1):125-54.
23. Hanney SR, Gonzalez-Block MA, Buxton MJ, Kogan M. The utilization of health research in policy-making: concepts, examples and methods of assessment. Health Res Policy Syst. 2003;1.
24. Kuruvilla S, Mays N, Walt G. Describing the impact of health services and policy research. J Health Serv Res Policy. 2007;12 Suppl 1:S1-23-31.
25. Burstein P. The impact of public opinion on public policy: A review and an agenda. Polit Res Quart. 2003;56(1):29-40.
26. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, et al. Reducing waste from incomplete or unusable reports of biomedical research. The Lancet. 2014;383(9913):267-76.
27. Pittman P. Health Services Research in 2020: Data and Methods Needs for the Future. Health Services Research. 2010;45(5 Pt 2):1431-41.
28. Wager E, Middleton P. Technical editing of research reports in biomedical journals. Cochrane Database Syst Rev. 2008(4):MR000002.
29. Fanelli D, Costas R, Lariviere V. Misconduct Policies, Academic Culture and Career Stage, Not Gender or Pressures to Publish, Affect Scientific Integrity. PLoS One. 2015;10(6):e0127556.
30. HSR-Europe. Health services research: helping tackle Europe's health care challenges. Policy Brief. Utrecht: NIVEL Netherlands Institute for Health Services Research, Technische Universität Berlin, National Board of Health - Danish Centre for Health Technology Assessment, Academic Medical Centre - University of Amsterdam, the London School of Hygiene and Tropical Medicine.; 2011.
31. Horton R. The hidden research paper. JAMA. 2002;287(21):2775-8.
32. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. BMJ. 2004;328(7454):1490.
33. de Groot AD. Methodologie. Grondslagen van onderzoek en denken in de gedragswetenschappen. Assen: Van Gorcum; 1994. p. 49.

SUPPLEMENTARY MATERIAL 1

Manuscript assessment & data extraction form (DEF)

Item		
1	**Assessor**	
1.1	Name	
1.2	Assessor role	
1.3	Assessor code	

2	**General information**	
2.1	Title of the study	
2.2	Journal	
2.3	Number of authors	
2.4	HSR (main) domain	
2.5	Involved institutions	
2.6	Funder(s) of the study	
2.7	Role of funder in the study	
2.8	Contribution of authors is stated	
2.9	Competing interests	
2.10	EQUATOR checklist available in additional materials	
2.11	Trial registration/protocol published	

3	**Introduction**		**Specify**	**Evaluation/ comments**
3.1	The objective(s) of the study are reported in the introduction			
3.2	The research question(s) are reported in the introduction			
3.3	The context of the study is explained			

4	**Methods**		**Specify**	**Evaluation/ comments**
4.1	Methodological approach			
4.2	Type of research			
4.3	Research design			
4.4	Data source is reported			
4.5	Selection of participants/sample is reported			
4.6	Non-response is reported			
4.7	Size of the study is reported			

4.8	Main outcome measure(s) are reported			
4.9	Secondary outcome measure(s) are reported			
4.10	Independent variable(s) are reported			
4.11	Description of quantitative and/or qualitative methods of analyses is reported			
4.12	Handling of missing data is reported			
4.13	Comparator is explained			

5	**Results**		**Specify**	**Evaluation/ comments**
5.1	Tables properly represent results			
5.2	Graphs properly represent results			
5.3	(Statistical) uncertainty is reported			

6	**Questionable messages and conclusions**	**QRP observed (0=no; 1= yes; -8 not applicable; -9=not assessable)**	**Evaluation/ comments (rationale for assessment of QRP)**	**Consulted project member (X= consulted for advice concerning methods, specifics about study, etc.)**
6.1	**Conclusions and key messages do not adequately reflect the objectives, design and actual findings**			
6.1.1	The title does not adequately reflect the main findings.			
6.1.2	The abstract does not adequately reflect the main findings.			
6.1.3	The conclusions in the abstract do not adequately reflect the conclusions in the main text.			
6.1.4	The objectives/research questions of the study are differently phrased in the introduction and the discussion.			
6.1.5	The outcome measure does not adequately reflect the objectives/ research questions of the study.			
6.1.6	The main results discussed in the discussion paragraph do not adequately address the original objectives/research questions as posed in the introduction.			

6.1.7	The order of presenting the results in de discussion is inconsistent with the ordering of the objectives/research questions as posed in the introduction.			
6.1.8	The conclusions do not adequately reflect the objectives of the study.			
6.1.9	The conclusions do not adequately reflect the findings as presented in the results paragraph.			
6.1.10	The outcome measure used does not allow the conclusions that are stated.			
6.1.11	The conclusion/discussion distracts from main outcomes by overstating the relevance of secondary outcomes.			
6.1.12	The conclusions are not supported by the results as presented in context of the referenced literature.			
6.1.13	Recommendations do not adequately reflect the results in context of the referenced literature.			
6.1.14	Implications for policy and practice do not adequately reflect the results in the context of the referenced literature.			
6.1.15	Lack of distinction between results and discussion. The results section contains elements of discussion and interpretation beyond the scope of explaining the results.			

6.2	**Main results are not or inadequately interpreted into the context of evidence**			
6.2.1	Supporting evidence is poorly documented.			
6.2.2	Contradicting evidence is poorly documented.			
6.2.3	Evidence is used inappropriately to support the findings (i.e. the argument is not supported by the actual message of the cited evidence). Will be measured as: Evidence seems to be used selectively to support the findings, given the title of the referenced evidence.			

6.2.4	The main source of evidence to support the results is based on the same underlying data.			

6.3	**Limitations are not adequately mentioned**			
6.3.1	Sources, direction and magnitude of bias are not or poorly discussed, or just listed without further discussion.			
6.3.2	The possible impact of the limitations on the results (i.e., magnitude and direction of any potential sources of bias) is not or poorly discussed.			

6.4	**Unjustified generalisations**			
6.4.1	The sampling methodology does not allow the type of generalization provided.			
6.4.2	Generalization of findings to populations not included in the original sample is not justified.			
6.4.3	Generalization of findings to time periods not included in the original study is not justified.			
6.4.4	Generalization of findings to geographical locations not included in the original study is not justified.			
6.4.5	Generalization of findings to settings/institutions not included in the original study is not justified.			

6.5	**Unjustified causation**			
6.5.1	Causative wording is used in the hypothesis/research question, although there is no theory supporting causation.			
6.5.2	A causal relationship is claimed, although the research design is not appropriate to determine causation (methods lack control of potential confounding or systematic bias).			
6.5.3	A causal relationship is claimed although potential sources of bias and their potential impact on the findings were not discussed.			
6.5.4	A potential causal relationship claimed in the discussion paragraph is not justified.			

6.6	**Effect size**			
6.6.1	The relevance of statistically significant results with small effect size is overstated.			
6.6.2	The possible clinical relevance of statistically nonsignificant results is not addressed.			
6.6.3	Non-significant results are discussed without addressing significance			

6.7	**Inappropriate use of language**			
6.7.1	Hyperboles and exaggerating adjectives are unjustifiably used (such as: key, groundbreaking, ideal, excellent, great, brilliant, extraordinary, impressive, completely, absolutely, entirely, everywhere, everything, nothing, beyond any doubt, definitely).			
6.7.2	Jargon, technical and complex language, that does not fit the journal audience, are used without properly explaining the meaning.			

7	**Miscellaneous**		
7.1	Overall qualitative evaluation of the study (e.g. quality, reporting style).		
7.2	Other comments.		

8	**Advice needed from second assessor**		
8.1	About the contents of the article		
8.2	Second assessment recommended		

Instructions per item

1	**Assessor**	
1.1	Name	
1.2	Assessor role	
1.3	Assessor code	

2	**General information**	**Instructions**
2.1	Title of the study	
2.2	Journal	
2.3	Number of authors	
2.4	HSR (main) domain	*Choose main discipline from list, add other disciplines in entry field*
2.5	Involved institutions	*List all-in*
2.6	Funder(s) of the study	
2.7	Role of funder in the study	*Copy funder declaration*
2.8	Contribution of authors is stated	
2.9	Competing interests	*Copy competing interest declaration*
2.10	EQUATOR checklist available in additional materials	
2.11	Trial registration/protocol published	*As mentioned in the article*

3	**Introduction**	
3.1	The objective(s) of the study are reported in the introduction	
3.2	The research question(s) are reported in the introduction	
3.3	The context of the study is explained	

4	**Methods**	
4.1	Methodological approach	
4.2	Type of research	
4.3	Research design	
4.4	Data source is reported	*e.g. registration, scientific or grey literature, survey data, interview data*
4.5	Selection of participants/sample is reported	*Selection of study enrolees also included case studies*
4.6	Non-response is reported	
4.7	Size of the study is reported	
4.8	Main outcome measure(s) are reported	

4.9	Secondary outcome measure(s) are reported	
4.10	Independent variable(s) are reported	
4.11	Description of quantitative and/or qualitative methods of analyses is reported	
4.12	Handling of missing data is reported	
4.13	Comparator is explained	

5	**Results**	
5.1	Tables properly represent results	*Tables give a reflection of actual results instead of cherry picking*
5.2	Graphs properly represent results	*Scaling is appropriate*
5.3	(Statistical) uncertainty is reported	*Confidence intervals are provided for the main results*

6	**Questionable messages and conclusions**	**Instructions**
6.1	**Conclusions and key messages do not adequately reflect the objectives, design and actual findings**	
6.1.1	The title does not adequately reflect the main findings.	*Title includes a quote or statement that does not accurately reflect/refers to the main findings, or deviates from the findings.*
6.1.2	The abstract does not adequately reflect the main findings.	*The abstracts contents deviate from / contradict with the main findings in the article text. Messy writing is not considered a QRP. Specifically for the conclusion in the abstract, causative wording misses: the conclusion in the abstract suggests causation, although the conclusions as discussed in the discussion paragraph report correlation. For instance, it is an unbalanced representation of the main results by focussing on secondary findings, while reducing the importance of the main findings, or reflects cherry-picking from the most conspicuous results. Or the stated results in the abstract in qualitative studies do not appear in the main text.*
6.1.3	The conclusions in the abstract do not adequately reflect the conclusions in the main text.	*The conclusions in the abstract are short-sighted compared to the actual conclusions in the main text. Conclusions can be stated in the discussion paragraph and/or the conclusion paragraph.*
6.1.4	The objectives/research questions of the study are differently phrased in the introduction and the discussion.	*When reporting objectives/research questions in the discussion. Different wording: does not need to include the exact wording, however the meaning/ connotation should be similar. Different ordering of objectives/research questions.*
6.1.5	The outcome measure does not adequately reflect the objectives/ research questions of the study.	*The objectives /research questions cannot be answered with the outcome measure that is studied*

6.1.6	The main results discussed in the discussion paragraph do not adequately address the original objectives/research questions as posed in the introduction.	*The research questions and/or objectives that were stated in the introduction section are not or only partly answered by the main results*
6.1.7	The order of presenting the results in de discussion is inconsistent with the ordering of the objectives/research questions as posed in the introduction.	*Not an actual QRP, but it does conflict with transparency in presenting the study's findings. If there's just one objective/research question, this item is not applicable (no structuring possible) and should be scored -8.*
6.1.8	The conclusions do not adequately reflect the objectives of the study.	*The objectives of the study are not met by the conclusions the study arrives at. Conclusions can be stated in the discussion paragraph and/or the conclusion paragraph. Either the study along the way shifted perspective, however no justification is provided. Or the write-up of the conclusions is flawed. Framing conclusion as extension to the discussion is not a QRP (undesirable, however beyond the scope if this indicator).*
6.1.9	The conclusions do not adequately reflect the findings as presented in the results paragraph.	*The conclusions deviate from the main findings. Conclusions can be stated in the discussion paragraph and/or the conclusion paragraph.* *The conclusion section does often not contain actual conclusions. The actual conclusion is often presented in the discussion section. Hence, conclusions in the discussion section are considered conclusions as well. Concluding statements will be marked, those statements that are only used to frame results (emphasizing importance of the study) are not considered conclusions. Key messages (in a box as separate section in some journals) are also considered conclusions.* *For instance, it is an unbalanced representation of the main results by focussing on secondary findings, while reducing the importance of the main findings, or reflects cherry-picking from the most conspicuous results.* *If new results are presented in the discussion section, then this is a QRP. (Assessors should not recalculate results)*
6.1.10	The outcome measure used does not allow the conclusions that are stated.	*For instance: the conclusions are about the quality of the health care system, whereas the outcome measure was 'satisfaction with home-care for elderly'*
6.1.11	The conclusion/discussion distracts from main outcomes by overstating the relevance of secondary outcomes.	*The main outcomes are ignored or their importance reduced, while favouring secondary outcomes. Most space is taken by discussing these secondary outcomes.*
6.1.12	The conclusions are not supported by the results as presented in context of the referenced literature.	*If the conclusion is not based on the results, but only on referenced literature, then this is noted as QRP (as aligns with 6.1.9). The extent of the conclusions is broader/more far fetching than the findings of the study, backed-up by discussed literature, justify. Conclusions can be stated in the discussion paragraph and/or the conclusion paragraph. For instance, a relationship between IV and DV is exaggerated. Conclusions cannot be stated based on referenced literature alone, main results are the fundament for the conclusions, that may be extended based on referenced literature.*

6.1.13	Recommendations do not adequately reflect the results in context of the referenced literature.	*Recommendations: what can/should be done with the study's findings? Recommendations are based on the results from the study, not only on the referenced literature. The extent of the recommendations is broader/more far fetching than the findings of the study, backed-up by discussed literature, justify. For instance, a relationship between IV and DV is exaggerated. QRP if no justification for the suggested recommendation is provided. QRP if no recommendation is provided.*
6.1.14	Implications for policy and practice do not adequately reflect the results in the context of the referenced literature.	*Implications: what are the consequences for policy and practice if the recommendations are followed-up? What would happen if the recommendations are carried out. (e.g. recommendations = implement the intervention in this setting, implication = the outcomes may improve by this much.) QRP if no justification for suggested implication is provided, QRP if no implication is provided. Originally: implications for policy and practice are poorly mentioned. Instruction: implications for practise and policy are well-balanced and give actual meaning to the findings of the study in context of practice and/or policy.*
6.1.15	Lack of distinction between results and discussion. The results section contains elements of discussion and interpretation beyond the scope of explaining the results.	*Applicable to all designs. Pilot included qualitative study, but also applies to quantitative studies. Results are placed in the context of literature beyond the theoretical model of the study.*

6.2	**Main results are not or inadequately interpreted into the context of evidence**	
6.2.1	Supporting evidence is poorly documented.	*Only limited evidence to support the main results is provided and only superficially discussed. No thorough reflection of the findings in perspective of supporting evidence.*
6.2.2	Contradicting evidence is poorly documented.	*Only limited evidence to oppose against the main results is provided and only superficially discussed. No thorough reflection of the findings in perspective of contradicting evidence.*
6.2.3	Evidence is used inappropriately to support the findings (i.e. the argument is not supported by the actual message of the cited evidence). Will be measured as: Evidence seems to be used selectively to support the findings, given the title of the referenced evidence.	*State inappropriately cited references, and explain why inappropriate: the evidence ascribed to the reference deviates from what could be assumed based on the title of the reference. Includes supporting results through self-citation (without further explanation of self-citation). Self-citation is not a QRP if clearly stated "in an earlier study we found…" If no references are used to support the results (QRP 6.2.1/2), then this is no QRP (QRP is avoided by not using literature), thus assessment is not possible and should be scored -9.*
6.2.4	The main source of evidence to support the results is based on the same underlying data.	*Most supporting evidence is grounded in the same data source as was used for the reviewed study (not necessarily self-citing), inducing circularity in argumentation.*

6.3	**Limitations are not adequately mentioned**	
6.3.1	Sources, direction and magnitude of bias are not or poorly discussed, or just listed without further discussion.	*Are the (relevant) limitations mentioned? The implications of the study design, methodology, sampling, context, etc. for risk of biasing study findings are not thoroughly discussed.*
6.3.2	The possible impact of the limitations on the results (i.e., magnitude and direction of any potential sources of bias) is not or poorly discussed.	*Is the impact of limitations discussed (if no limitations are mentioned then this is considered a QRP). The extent to which potential risks of bias affect the interpretation of the findings is not thoroughly discussed.*

6.4	**Unjustified generalisations**	
6.4.1	The sampling methodology does not allow the type of generalization provided.	*The sample is too specific, small, or flawed (for instance by attrition, selection bias) for the generalization that is made.*
6.4.2	Generalization of findings to populations not included in the original sample is not justified.	*The included sample is too specific, small or flawed (for instance by attrition, selection bias) and no or inadequate evidence is provided to support the generalization that is made. Population does not include geographical location (this is a separate QRP). Population includes population characteristics such as gender, ethnicity, age, etc.*
6.4.3	Generalization of findings to time periods not included in the original study is not justified.	*The characteristics of the included time period are too specific (for instance in election period, affecting the policy that was studied) and no or inadequate evidence is provided to support the generalization that is made*
6.4.4	Generalization of findings to geographical locations not included in the original study is not justified.	*The characteristics of the included geographical location(s) are too specific to generalise to other geographical locations (for instance very urbanised area to rural setting) and no or inadequate evidence is provided to support the generalization that is made*
6.4.5	Generalization of findings to settings/institutions not included in the original study is not justified.	*The characteristics of the included institutions are too specific to generalise to other institutions (for instance hospital regulations to nursing homes) and no or inadequate evidence is provided to support the generalization that is made*

6.5	**Unjustified causation**	
6.5.1	Causative wording is used in the hypothesis/research question, although there is no theory supporting causation.	*Quantitative: hypothesis is not justified/allowed since there's no theory to support a causal relationship*
6.5.2	A causal relationship is claimed, although the research design is not appropriate to determine causation (methods lack control of potential confounding or systematic bias).	*No causation based on the results of the present study may be assumed if no RCT is conducted... (or longitudinal cohort?)*

6.5.3	A causal relationship is claimed although potential sources of bias and their potential impact on the findings were not discussed.	*No or inadequate discussion is included concerning the impact of potential sources of bias on the possible causation that was found in the results*
6.5.4	A potential causal relationship claimed in the discussion paragraph is not justified.	*When a causal relation may not be assumed solely based on the study's findings, no or inadequate supporting and contradicting evidence is used to discuss the possible causation that was found in the results.*
6.6	**Effect size**	
6.6.1	The relevance of statistically significant results with small effect size is overstated.	*Importance of findings is exaggerated. Although (some) results are statistically significant, the clinical/practical relevance is minor due to small effect size/causation is unlikely.*
6.6.2	The possible clinical relevance of statistically nonsignificant results is not addressed.	*Importance of findings is dismissed, since no statistical significance was reached. Although the findings reflect likely causation and non-significance was likely due to lack of power.*
6.6.3	Non-significant results are discussed without addressing significance	*Results are discussed as if they were significant, without addressing they are not, or what the uncertainty is.*
6.7	**Inappropriate use of language**	
6.7.1	Hyperboles and exaggerating adjectives are unjustifiably used (such as: key, groundbreaking, ideal, excellent, great, brilliant, extraordinary, impressive, completely, absolutely, entirely, everywhere, everything, nothing, beyond any doubt, definitely).	*The use of adjectives that exaggerate the relevance of the findings, conclusions and messages. Not actually counting adjectives, if one hyperbole is used and attracted the attention. Hyperbolic adjective use per se is no QRP, only in relation to results/conclusions, to exaggerate the study's findings.*
6.7.2	Jargon, technical and complex language, that does not fit the journal audience, are used without properly explaining the meaning.	*The journal audience is not properly addressed by the language used. Language use seems to be overly complex to impress or distract the reader.*

7	**Miscellaneous**	
7.1	Overall qualitative evaluation of the study (e.g. quality, reporting style).	*If a certain aspect impacts the answer to multiple questions, specify in "other comments". E.g. if the discussion section does not contain main results, then this item cannot be assessed.*
7.2	Other comments.	

8	**Advice needed from second assessor**	
8.1	About the contents of the article	*What advice is needed, state question.*
8.2	Second assessment recommended	*First assessor doubts about assessment and requests second opinion.*

SUPPLEMENTARY MATERIAL 2

Additional information to the methods of the development of the definition and measurement instrument for "questionable research practices in the reporting of messages and conclusions in scientific health services research publications"

This document describes the methods used to develop a definition of questionable research practices (QRPs) in the reporting of messages and conclusions, and to construct a measurement instrument that allows for the identification of questionable research practices in the reporting of messages and conclusions in Health Services Research (HSR).

Methodology

Methods included an explorative review of definitions in literature, a consultation meeting with the project group, institution/department leaders of Dutch HSR institutions and project advisors (n=13), semi-structured interviews with 13 HSR institutes (n=19) and an expert consultation (n=5).

Setting

13 HSR groups, departments, or institutions (hereafter referred to as "HSR institutions") in the Netherlands, including both academic and non-academic institutions participated in this study. These institutions all agreed to participate in an effort to assure the overall quality of HSR publications in the Netherlands.

Literature review

First, a literature review was conducted searching for existing definitions of questionable research practices in the reporting of conclusions and messages, and operationalisations of QRPs. Search terms included in different order and combination: 'questionable research practices', 'spin', 'over interpretation', 'discordant conclusions', 'QRPs', 'outcome reporting bias', 'questionable conclusions' and 'responsible conclusions'. Documents were included if they described methods to measure questionable research practices in scientific publication, or provided definitions of the above key terms. Referred documents that fit the criteria were also included in the review.

After identifying the main literature that suited our aim, we came to a preliminary definition of QRPs based on Boutron 2010, Ochobo 2013, and Horton 1995 (1-3).

An extensive list of possible types of QRPs in the reporting of messages and conclusions was developed, based on the EQUATOR checklists(4) and instruments from previous studies. For example, instruments for identifying ‘spin’, reporting of qualitative research and other QRPs such as (3, 5-7). Spin in this context refers to “a way to distort science reporting without actually lying”)

Consultation meeting

Second, we presented the preliminary QRP definition and the first draft of items referring to QRPs (see page 3) during a consultation meeting of participating HSR institutions on 6 June 2017. The meeting lasted three hours, during which the research project and the preliminary definition and draft of QRP items was discussed. Representatives of the participating HSR institutions (n=7), project advisors (n=2) and project group members (n=4) attended the meeting. The attendees discussed their thoughts about the definition and its operationalisation. Detailed notes from this meeting were summarized and shared with the representatives of all participating institutes (including those who did not attend).

The central conclusion of the meeting was to focus on the ‘measurability’ of the QRPs. An important consideration in developing the instrument for the assessment of scientific publication is to focus on the possibility to measure the QRPs. Therefore, the focus should be on QRPs that can be quantified. These should be distinguished from QRPs that, although possibly important, are not quantifiable.

Semi-structured interviews

Third, we conducted fourteen semi-structured interviews with nineteen leaders/representatives of the thirteen HSR institutions. These representatives had to have a clear overview of the process of reporting research in their institute. One of the institutions was represented by two separate departments, hence two representatives were separately interviewed. Three interviews were conducted with both the institute leader and a second representative. One of the interviews included three representatives of an institution. The aim of the interviews was to discuss our draft of QRP items and identify additional measurable QRPs in the reporting of messages and conclusions in HSR, explore potential causes of QRPs in messages and conclusions, and to discuss experiences of the institute leaders with these QRPs. A semi-structured interview guide was developed by the project team. During the interview, we presented the interviewees with a draft of QRP items. The draft list was iteratively adjusted, i.e. after each interview we drafted a new version including the findings of the previous interviews.

Interviewees were approached through e-mail to schedule an appointment. Two researchers conducted the interviews of which thirteen took place at the participating institutions and one interview took place in a public space. During the first interview, both researchers were present to align the approach. The remainder of the interviews were equally divided between them. The interviews lasted one hour. In concordance with ethical guidelines, the goal of the interview was explained at the start of the interview and permission to audio-record the interview was obtained.

With the support of the recordings, a report was written and shared with the interviewees for validation. All interviewees confirmed the reports, after mostly minor edits to the report. From the interview reports, we drew up a new draft of the list of QRP items. In the research group, we specifically paid attention to correct wording of the QRPs.

Expert consultation

Fourth, ten leading international health services researchers were asked to provide feedback on this list of QRP items. These HSR experts were invited through e-mail in which we explained the aim of the study, and included the definition of QRPs and the list of QRP items. Five experts provided their comments to the items. Five experts did not respond after a reminder, or indicated not having time to review the QRP items. Feedback was summarized, and comments were used to adapt the QRP definition and list of QRP items.

Measurement instrument

We developed the measurement instrument in Excel format by taking items from earlier developed checklists (EQUATOR and COREQ) and the list of QRPs. The measurement instrument was completed after a final consensus meeting of the research group. The measurement instrument exists of three sections: 1) bibliographic information of the publication (eg. funder, journal, number of authors), 2) basic methodological information (eg. included population, analyses method) and 3) possible QRPs in messages and conclusions. A pilot was conducted to assess the feasibility and usability of the instrument. In the pilot, two project members independently assessed five international HSR publications to identify modifications needed to improve the items in the instrument, and to align the interpretation of the items. The project group discussed the proposed modifications, resulting in the final version: the data extraction form (see supplementary material 1.)

SUPPLEMENTARY MATERIAL 3

Table S3.1 | Occurrence of QRPs in the reporting of messages and conclusions in HSR publications (n=116), ordered from most frequently to least frequently occurring (%).

Questionable research practices (QRPs) in reporting messages and conclusions	% publications with QRP	% publications without QRP	% publications for which QRP not assessable
Implications for policy and practice do not adequately reflect the results in the context of the referenced literature.	**69·0	31·1	0·0
Recommendations do not adequately reflect the results in the context of the referenced literature.	***65·5	34·5	0·0
Contradicting evidence is poorly documented.	63·8	36·2	0·0
Conclusions do not adequately reflect the findings as presented in the results section.	46·6	51·7	1·7
Possible impact of the limitations on the results is not or poorly discussed.	44·0	56·0	0·0
Conclusions are not supported by the results as presented in the context of the referenced literature.	43·1	54·3	2·6
The conclusions do not adequately reflect the objectives of the study.	35·3	61·2	3·4
Supporting evidence is poorly documented.	31·9	68·1	0·0
Sources. direction and magnitude of bias are not or poorly discussed. or just listed without further discussion.	27·6	72·4	0·0
The conclusions in the abstract do not adequately reflect the conclusions in the main text.	22·4	75·0	2·6
The main results discussed in the discussion paragraph do not adequately address the original objectives/research questions as posed in the introduction.	20·7	75·9	3·4
The outcome measure used does not allow the conclusions that are stated. *	18·1	81·9	0·0
Lack of distinction between results and discussion. The results section contains elements of discussion and interpretation beyond the scope of explaining the results.	17·2	82·8	0·0
The sampling methodology does not allow the type of generalization provided.	15·5	84·5	0·0

Table S3.1 | Continued

Questionable research practices (QRPs) in reporting messages and conclusions	% publications with QRP	% publications without QRP	% publications for which QRP not assessable
The objectives/research questions of the study are differently phrased in the introduction and the discussion.	14·7	36·2	49·1
The order of presenting the results in de discussion is inconsistent with the ordering of the objectives/research questions as posed in the introduction.	14·7	75·0	10·3
Hyperboles and exaggerating adjectives are unjustifiably used	12·1	87·9	0·0
The title does not adequately reflect the main findings.	11·2	88·8	0·0
The abstract does not adequately reflect the main findings.	10·3	89·7	0·0
A potential causal relationship claimed in the discussion paragraph is not justified.	10·3	89·7	0·0
The outcome measure does not adequately reflect the objectives/research questions of the study. *	9·6	90·4	0·0
A causal relationship is claimed. although the research design is not appropriate to determine causation.	9·6	90·4	0·0
The relevance of statistically significant results with small effect size is overstated.*	9·6	90·4	0·0
Generalising findings to settings/ institutions not included in the original study is not justified.	9·5	89·7	1·0
The conclusion/discussion distracts from main outcomes by overstating the relevance of secondary outcomes. *	8·4	91·6	0·0
Non-significant results are discussed without addressing significance.	8·4	91·6	0·0
Generalising findings to geographical locations not included in the original study is not justified.	6·0	94·0	0·0
Evidence is used inappropriately to support the findings.	5·2	94·9	0·0
A causal relationship is claimed although potential sources of bias and their potential impact on the findings were not discussed. *	3·6	96·4	0·0

Table S3.1 | Continued

Questionable research practices (QRPs) in reporting messages and conclusions	% publications with QRP	% publications without QRP	% publications for which QRP not assessable
Jargon. technical and complex language. that does not fit the journal audience. are used without properly explaining the meaning.	3·4	96·6	0·0
The main source of evidence for supporting the results is based on the same underlying data.	2·6	96·6	0·9
Generalising findings to populations not included in the original sample is not justified.	2·6	97·4	0·0
Causative wording is used in the hypothesis/research question, although there is no theory to support causation. *	2·4	97·6	0·0
Possible clinical relevance of statistically non-significant results is not addressed. *	2·4	97·6	0·0
Generalising findings to time periods not included in the original study is not justified.	0·0	100·0	0·0

** QRPs only applicable to quantitative research-based publications (n=83)*
*** 50 ·0% of publications did not mention implications for policy or practice.*
**** 34·5% of publications did not mention recommendations for policy or practice.*

6

Reporting Health Services Research to a broader public: An exploration of inconsistencies and reporting inadequacies in societal publications

AUTHORS:
Reinie G. Gerrits, Michael J. van den Berg, Anton E. Kunst, Niek S. Klazinga, Dionne S. Kringos

Submitted for publication

ABSTRACT

Researchers are increasingly encouraged to translate their scientific work into societal publications to reach broader audiences. This study investigates possible inconsistencies and reporting inadequacies in non-scientific societal publications based on published health services research (HSR). We conducted a directed qualitative content analysis of 43 scientific HSR publications and 156 societal publications derived from them. We investigated whether inconsistencies in societal publications were less likely when the first scientific author was involved. 60.3% of the 156 societal publications (associated with 74.4% of the scientific publications) contained messages that were inconsistent with the scientific work. Reporting inadequacies in 51.2% of the scientific publications were replicated in associated societal publications (28.9%). The likelihood of inconsistencies between scientific and societal publications did not differ when the latter explicitly involved the first scientific author, $\chi^2(2, n=43) = 3.2$, $p=.07$; were published on the institute's or funder's website, $\chi^2(2, n=29) = 0.41$, $p=.52$; published with no involvement of the scientific author, $\chi^2(2, n=84) = 1.2$, $p=.267$.

To improve societal publications on scientific research, one should examine not only how consistency with scientific research publications can be achieved, but also how to prevent scientific reporting inadequacies from getting replicated in societal publications.

BACKGROUND

In academia, scientific research publications are an important source of knowledge, as well as a means of research dissemination (1). Outside the research community, however, most people take note of research findings through non-scientific, societal publications such as press releases, newspapers, social media, internet postings or professional journals (2-4). The content of societal publications impacts the thinking, debates and decisions of the general public, as well as those of patients, health professionals and policymakers (4-6). Consequently, researchers who publish a scientific paper are increasingly incentivised to 'translate' their findings into a corresponding societal publication, in order to reach broader, often non-academic audiences (7).

By necessity, the authors of societal publications simplify scientific messages and conclusions for their lay target group (8). Although this can be done in a responsible manner, it does present a risk for misrepresentation and misinterpretation of the research findings (9). Previous studies on biomedical publications concluded that unjustified causal claims are introduced in 20% to 33% of press releases, and that 40% of news articles give more explicit health advice to the readers than was expressed in the underlying scientific publication (10-14).

In health services research (HSR), less is known about the potential misrepresentation of evidence in societal publications. HSR aims to provide usable evidence for policy and for management of health and health care (5). This practice-oriented ambition amplifies the importance of accuracy in all messages and conclusions relayed in societal publications (15).

Researchers are often expected to have a societal impact beyond their scientific impact. Funders of HSR increasingly demand strategies to achieve a societal impact. Methods for measuring impact are being developed and refined (16-18). Researchers, however, may lack the experience or capability to write responsible societal publications that accurately reflect their scientific findings (10, 19). Some previous research has concluded that a researcher's involvement is not associated with better societal publications (10). Researchers may have difficulty working with journalists, or they may lack the ability to explain their findings in simple terms (19). Moreover, fellow researchers may take a critical view of colleagues who invest considerable time in media attention, and thus discourage them to put significant efforts in writing societal publications (20).

Messages and conclusions are not always well reported in the scientific publications either. In a previous assessment of peer-reviewed HSR publications written in the Netherlands for an international academic audience, we found per publication a median of 6 out of 35 possible ‘questionable research practices’ (QRPs) in the reporting of messages and conclusions (21). In the current study, these QRPs will be called ‘reporting inadequacies’. They include conclusions that are insufficiently supported by the research results, recommendations that are not justified and limitations that are inadequately explained (21). Even if a researcher tries to avoid inconsistencies in a subsequent societal publication, such reporting inadequacies in the original work may well find their way to a broader audience. As the scientific publication is used as the standard, reporting inadequacies will likely be copied or ‘replicated’ to societal publications.

Given the potential impact of societal publications on policy and practice, knowledge of responsible reporting in societal publications, and how researchers can achieve it, is important for the HSR community (22). Such knowledge is currently inadequate (10, 14). Whereas the previous studies in the field of biomedicine focused largely on press releases and newspapers, broader insights are needed into the full scope of societal HSR publications, including information sources such as fact sheets, web pages and articles in professional journals.

The aims of this study are to explore

1. whether societal publications on HSR are consistent with the messages reported in the underlying research papers
2. whether reporting inadequacies in scientific HSR publications are replicated in societal publications
3. whether fewer inconsistencies occur in societal publications if the first scientific author is involved in writing them.

METHODS

In a collaboration funded by the Netherlands Organisation for Health Research and Development (ZonMw), thirteen Dutch academic and non-academic HSR institutions (see Acknowledgements section for the listing) took part in several studies designed to promote responsible reporting. The present study builds on the results of a previous study that identified reporting inadequacies in scientific publications (21).

To investigate inadequacies in research reporting and inconsistencies between scientific and societal publications, we conducted a directed qualitative content analysis of 43 scientific HSR publications and 156 related societal publications that derived from them.

Subsample of scientific publications

We based our selection of scientific publications on a random sample of 116 such publications authored in 2016 by researchers from the thirteen participating HSR institutions. Our previous study assessed those publications for inadequacies in the reporting of messages and conclusions, using a checklist of 35 possible inadequacies. Each inadequacy was recorded on an assessment form.

A description of the sampling of these publications has been published elsewhere.(21) In short, complete publications lists were obtained from all institutions, from which a total of 717 scientific HSR publications were identified, applying commonly used definitions of HSR from Juttman and collegues (23) and Lohr and Steinwachs.(24) A sample of 116 publications was assessed.

The current study confined itself to scientific writings that had one or more associated societal publications. We sampled until we identified 23 scientific publications with a relatively high number of reporting inadequacies and 23 with a relatively low number, based on the median of inadequacies per publication ('high' being more than 6 and 'low' fewer than 6).

Sample of societal publications

Societal publications corresponding to scientific publications were identified through a systematic internet search. We included societal publications that (1) were in the public domain and (2) contained messages on the same research as the corresponding scientific publication (including statements on the results, conclusions, discussion, recommendations or implications).

For each scientific publication, a variety of internet sources were consulted, following a systematic search strategy. We searched or consulted (1) specific institute websites, funders' websites and Altmetrics; (2) document databases of Dutch government and parliament (including *www.rijksoverheid.nl*, *www.tweedekamer.nl*, and *https://zoek.officielebekendmakingen.nl/*); (3) databases of Dutch popular science periodicals (*https://www.skipr.nl/zoeken?q*, *https://www.medischcontact.nl)* and a periodical aimed at medical professionals *https://www.ntvg.nl/zoeken*); (4) an existing database of Dutch newspaper articles (*www.lexisnexis.nl*); (5)

public social media platforms (LinkedIn and Twitter) of the authors and the institutes; (6) the Google search engine, to identify publications from further sources. In the Google search, we entered search terms (see next section) and followed all links provided in the first 30 results, as we did not expect to find relevant societal publications beyond that ranking. To ensure that earlier searches did not affect the Google search, our browser history data, including cookies, were deleted beforehand. All internet sources were accessed in the period of August 2018.

Search terms and filters

For each scientific publication, specific search terms were derived from Dutch translations of key terms in the title and abstract; also included were the name of the first author's institution, the authors' names and the funder(s). Any new key terms found during the search were added.

Search strings were used if the database enabled the use of logical operators. Because results of a study may be reported prior to the appearance of the scientific publication, we included societal publications appearing up to two years beforehand and one year afterwards (presuming that all societal publications would appear within a year). All search terms were discussed and approved by two members of the project team (RG and NK).

Analyses of messages in societal publications

Societal publications were analysed using a directed content analysis approach. (25) First, we identified distinct messages and conclusions in the societal publications that related to the corresponding scientific publication. Messages could be a single sentence or a section of the text elaborating on the same topic; a single research result or a concluding statement was marked as a distinct message. Multiple messages might be identified in a single societal publication.

Second, we assessed whether the message in the societal publication was consistent with that in the corresponding scientific work. A message was considered consistent if it conveyed the same meaning as the scientific results, discussion or conclusion and if no changes, additions or subtractions had been made with respect to the content of the scientific assertion. An initial coding scheme based on the possible inconsistencies was prepared, informed by other checklists for public reporting. (12, 26) To discover other types of inconsistencies not included in those checklists, we iteratively improved the coding scheme during the first stages of coding, adding new aspects that emerged during the coding.

Third, we determined whether a message in the societal publication replicated a reporting inadequacy in the scientific publication. Messages in the societal writings were compared to any reporting inadequacies recorded during the previous assessment of the corresponding scientific paper (21). A message that identically reproduced the reporting inadequacy was marked as a 'replicated reporting inadequacy'.

Finally, for each societal publication we gauged the likelihood of the first scientific author's involvement ('named as author', 'published on institute or funder web page' or 'no involvement').

One coder (RG) performed the analysis. The identification of the messages and the coding method of the first ten publications were checked by project members (NK, DK and MB) and thoroughly discussed until the coding method and scheme had been agreed. To ensure consistency of analyses, we had 10 per cent (*n* = 16) of the analysed societal publications randomly checked by DK. Prompted by the check, we decided to revisit the final 15 societal publications to improve possible inconsistencies with earlier codes, and to correct one identified replicated reporting inadequacy. Analyses were conducted in MAXQDA.

Statistical analyses

Chi-square tests of independence were calculated to compare the frequencies of inconsistencies in societal publications (1) authored by the first scientific author, (2) published on the institute's or funder's web page or (3) published with no involvement of the scientific author. Three tests were performed, each comparing one category to the other two categories combined. A societal publication was deemed 'inconsistent' if at least one message in it was identified as inconsistent.

RESULTS

Characteristics of analysed publications

We conducted the structured internet search until we identified 46 scientific publications (23 with high and 23 with low inadequacies) that had associated societal publications. We identified the included 46 publications after examining 84 scientific publications in our sample (46 with the highest number of inadequacies and 38 with the lowest number of inadequacies). We examined 188 societal publications obtained in our internet search and excluded 32 of them from further analysis because they described methodology or study results only. That left three further scientific publications without corresponding societal ones, so that 43 scientific publications and 156 societal publications remained in the final sample. Table 1 shows the characteristics of the included societal publications.

Scientific publications had a maximum of 14 associated societal publications, a minimum of 1 and a median of 3. The majority (*n* = 27) had 1 to 3 societal publications, 11 had 4 to 6 societal publications, and 5 had more than 6. A total of 60.3% of the societal publications corresponded to a scientific publication with low reporting inadequacies; 39.7% were linked to a publication with high inadequacies.

Table 1 | Characteristics of the analysed societal publications

Type of societal publication	**Societal publications, *n* (%)**
News message	37 (23.7)
Institute web page	21 (13.5)
Magazine	20 (12.8)
Social media	19 (12.2)
Professional journal	19 (12.2)
Report	13 (8.3)
Thesis summary	9 (5.8)
Funder web page	8 (5.1)
Fact sheet	7 (4.5)
Video	2 (1.3)
PowerPoint slides	1 (0.6)
Linkage to scientific reporting inadequacies	**Societal publications, n (%)**
Linked to high (>6) reporting inadequacies in scientific publications (*n* = 20)	62 (39.7)
Linked to low (<6) reporting inadequacies in scientific publications (*n* = 23)	94 (60.3)
Total societal publications	**156 (100.0)**

Consistency of reporting between societal and scientific publications

In the 156 analysed societal publications, we identified 577 distinct messages, 342 (59.3%) of which were consistent with the corresponding message in the underlying scientific publication. Multiple types of inconsistencies were sometimes identified within a single societal publication, such as altered conclusions or differing interpretations of outcomes.

Inconsistencies in conclusions (64 societal publications, 41.0%)

The majority of inconsistencies in societal publications concerned the scientific research conclusions. Conclusions might be altered entirely (in 13 publications) or partially (in 14). Moreover, conclusions were presented that were not underpinned by the scientific results or discussion. Some publications gave interpretations of the study results that were not included in the scientific work (17 publications) or added strong rhetoric to conclusions that was inconsistent with the scientific verdict (e.g. implying that a problem was worse; 12 publications). Some societal publications put forward conclusions that could not have resulted from the scientific study (21 publications) or that were derived from the introduction (4 publications). Some societal publications even contradicted the conclusions from the corresponding scientific publications (10 societal publications).

Inconsistencies in results (38 societal publications, 24.4%)

In various societal publications, new results were introduced that were not reported in the corresponding scientific publication (22 societal publications). Results were reported in different combinations that changed the interpretation of the results (in 7 publications). Some quantitative results were altered with respect to the figures given in the scientific publication (e.g. differing exact percentages) or qualitative results were worded differently, causing the core meaning of the scientific findings to change (13 publications). Non-significant results were presented as significant (in 1 publication), such as referring to a 'lesser effect from this intervention', whereas no effect had been indicated or argued in the scientific publication.

Inconsistencies in recommendations (25 societal publications, 16.0%)

Recommendations differed from those made in scientific publications in three ways: (1) entirely new recommendations for policy or practice were put forward in the societal publication, whilst not mentioned in the scientific publication (in 21 societal publications); (2) relevant limitations of the recommendations given in the scientific publication were omitted in the societal publication (1 publication); (3) elements of recommendations given in societal publications were omitted in the scientific publication (6 societal publications).

Inconsistencies in the reporting of conditions in the conclusion (4 societal publications, 2.6%)

In four societal publications, conditions affecting the study conclusions were left out, although the scientific publication explicitly made the conclusions subject to those conditions (e.g. with an 'if' or 'when' statement as part of the conclusion).

Inconsistencies in the reporting of implications for policy and practice (4 societal publications, 2.6%)

Implications for policy and practice must be differentiated from recommendations: implications describe the importance of the findings for policy and practice, while recommendations are specific measures that could improve policy and practice. In four societal publications, implications for policy and practice were reported that were not mentioned in the scientific publication.

Inconsistencies in the reporting of causality (4 societal publications, 2.6%)

Four societal publications contained statements on potential causal relationships that were not mentioned in the scientific publication, and causality was implied without mention of mediating influences.

Inconsistencies in reporting generalisations (3 societal publications, 1.7%)

Three societal publications generalised findings beyond the setting described in the scientific publication – to a different time period or geographical location, as from an urban to a rural setting; to different population characteristics such as gender, ethnicity or age; or to settings or institutions not included in the research.

Objectives not included in the scientific publication (1 societal publication, 0.6%)

In one societal publication, a study objective was added and discussed that was not included in the scientific publication (nor in any related research project).

Replication of reporting inadequacies from scientific to societal publications

Reporting inadequacies found in 51.2% (*n* = 22) of the included scientific publications were reproduced in corresponding societal publications. From our checklists of inadequacies in scientific reporting, we identified nine types of inadequacies that were replicated in societal publications:

- 'Conclusions do not adequately reflect the findings as presented in the results section' (from 23.3%, *n* = 10, of the scientific publications)
- 'Recommendations do not adequately reflect the results in the context of the referenced literature' (26.3%, *n* = 7)
- 'The title does not adequately reflect the main findings' (9.3%, *n* = 4); that is, the inadequate title of the scientific publication was replicated in a societal publication.
- 'The sampling methodology does not allow the type of generalisation provided' (7%, *n* = 3)
- 'The conclusions in the abstract do not adequately reflect the conclusions in the main text [of the scientific publication]' (4.7%, *n* = 2); that is, inadequately reported conclusions from the abstract were replicated in a societal publication.
- 'A potential causal relationship claimed in the discussion paragraph is not justified' (4.7%, *n* = 2)
- 'Implications for policy and practice do not adequately reflect the results in the context of the referenced literature' (2.3%, *n* = 1)
- 'The abstract does not adequately reflect the main findings' (2.3%, *n* = 1)
- 'Generalising findings to geographical locations not included in the original study is not justified' (2.3%, *n* = 1)

The role of the first scientific author in inconsistencies appearing in societal publications

From our sample of 43 scientific publications, 26 first authors were named as authors of a societal publication (60.5%). Some 34 scientific publications were linked to at least one societal publication that did not explicitly state involvement of the first author (79.1%). Research from 20 scientific publications was summarised on the website of a research institute or funder without explicit mention of the involvement of the author (46.5%).

Chi-square tests of independence were calculated comparing the frequencies of inconsistencies in societal publications (1) authored by the first author of the scientific publication, (2) published on the institute's or funder's website, and (3) published elsewhere without explicit involvement of the scientific author. No associations were found between the number of inconsistencies in societal publications and any of those three conditions (table 2), although a trend to fewer inconsistencies was observed when the first author wrote the societal publication, χ^2 (2, *n* = 43) = 3.2, *p* = .07.

Table 2 | Chi-square test of independence on the role of the first scientific author in the occurrence of inconsistencies between a societal and a scientific publication (N = 156)

Type of societal publication	Author involvement, *n* (%)	At least one inconsistency found, *n*		Chi-square probabilities
		Yes	No	
No involvement of scientific author	84 (53.9)	54	30	χ^2 1.2 (p = .267)
Authored by first scientific author	43 (27.6)	21	22	χ^2 3.2 (p = .07)
Published on institute or funder website	29 (18.6)	19	10	χ^2 0.41 (p = .52)

Consistencies and replicated reporting inadequacies across scientific publications

Following our analyses, the sample of scientific publications (N = 43) could be broken down into four unique groups in relation to the associated societal publications:

- All corresponding societal publications were consistent and did not replicate any reporting inadequacies (n = 7 scientific publications).
- Corresponding societal publications replicated reporting inadequacies, but were fully consistent with the scientific publication (n = 4).
- Corresponding societal publications were inconsistent, but did not replicate reporting inadequacies (n = 15).
- Corresponding societal publications were inconsistent with the scientific publication and replicated reporting inadequacies (n = 17).

DISCUSSION

The aims of this study were to explore (1) whether societal publications on health services research are consistent with the messages communicated in the original scientific research paper, (2) whether apparent reporting inadequacies in scientific HSR publications are replicated in societal publications, and (3) whether fewer inconsistencies occur in societal publications if they are authored by the first author of the scientific work. 60.3% of the 156 societal publications (associated with 74.4% of the scientific publications) contained messages that were inconsistent with the scientific work. Reporting inadequacies in 51.2% (n = 22) of the scientific publications were replicated in corresponding societal publications (n = 45, 28.9%). The involvement of the first author was not associated with more consistent societal publications, although a slight trend was indicated.

Our results indicate that, as previously shown for biomedicine, the field of HSR faces issues with (mis)representation and (mis)interpretation of the research findings, as reported in societal publications (10-14). Such issues arise not only in news articles or press releases, but also in societal publications such as professional journal articles aimed directly at policy and practice.

Limitations

As our coding scheme was not specifically designed to identify causality, we have likely underestimated the occurrence of causal claims. The coding schemes used in previous studies, though very extensive, would not have been adequate for detecting many types of inconsistencies, such as rhetorical formulations of conclusions or diverging interpretations of results, as we have done in this study. In addition, the existing coding schemes would not have been suitable for HSR, as different types of systematic research were addressed here, including qualitative and mixed methods studies, and different types of societal publications were included in our analyses, such as tweets and fact sheets.

We analysed whether assertions in a societal publication were consistent with those in the corresponding scientific publication. We did not assess omitted messages; that is, we did not identify scientific reporting inadequacies attributable to the absence of common elements such as limitations, recommendations or contradictory evidence. Consequently, we also did not take a normative stand on whether those items should have been included in a societal publication. Such would not have been feasible considering the variety of societal publications studied, ranging from tweets to professional journals.

The numbers of associated societal publications were not equally distributed over the included scientific publications; one scientific author of multiple societal publications could have skewed our results. We therefore recommend further research on the roles of individual researchers in writing responsible societal publications.

Our sample of scientific publications was small and insufficiently wide-ranging to determine the prevalence of reporting inadequacies and inconsistencies across the field of HSR internationally.

Interpretation

Our results indicate that most societal publications contain some inconsistencies or replicated reporting inadequacies. Inconsistencies are not necessarily negative, as they may correct an inadequacy in the scientific publication. Moreover, reporting inadequacies we identified in this study were not necessarily 'bad'. There is no straightforward rule for what is allowed in terms of rhetorical wordings or simplifications of scientific results in either scientific or societal publications. However, the current discussion on public reporting is focused too narrowly on exaggeration and causality (10, 27, 28). There is very little debate on questions such as whether conclusions and recommendations are adequately reported in scientific literature, the extent to which messages in societal publications may justifiably be simplified, how much detail needs to be provided, and whether a researcher or journalist may add interpretations in societal publications that would not be accepted in scientific literature.

Reporting inadequacies in scientific publications commonly get replicated in societal publications. Most frequently this involves inadequately reported conclusions, policy and practice recommendations, and titles. It is therefore insufficient to focus merely on preventing inconsistencies in societal publications. We recommend that future studies that assess quality in societal publications should extend their research questions to analyse this interplay between the reporting in scientific publications and societal publications.

While we found a trend whereby first authors appeared to write more accurate societal publications, no substantial differences emerged overall between societal publications produced by research institutes or funders and ones written by outsiders. A stronger relation between the involvement of researchers in writing societal publications and consistency with their scientific publications may be desirable. Media pressures, relationships with funders, and journal demands may cause researchers to consciously or unconsciously introduce reporting inadequacies into a scientific publication (29-31).

Implications and recommendations for policy and practice

Researchers, research institutes and journalists should be attentive to the effects that the rewriting of research results and conclusions in societal publications might have on policy and practice. Additionally, researchers should be aware that reporting inadequacies in their scientific publications may get replicated in societal publications and subsequently affect policy and practice. Routines such as peer feedback in the final stages of publication could prevent such reporting inadequacies from occurring in scientific publications. Further training and time dedicated to societal reporting and to communicating about scientific work in lay language would better equip researchers to take active roles in the writing of societal publications.

CONCLUSION

To improve societal publications on health services research, we should examine not only how consistency with scientific research publications can be achieved, but also how to prevent scientific reporting inadequacies from being replicated in societal publications. HSR institutions, funders, and scientific and societal publication platforms should invest in a supportive publication culture in order to further incentivise the responsible and skilled involvement of researchers in writing both scientific and societal publications.

REFERENCES

1. Weingart P. Science and the media. Res Policy. 1998;27(8):869-79.
2. Ritter A. How do drug policy makers access research evidence? Int J Drug Policy. 2009;20(1):70-5.
3. Zardo P, Collie A. Type, frequency and purpose of information used to inform public health policy and program decision-making. Bmc Public Health. 2015;15.
4. Gerrits RG, van den Berg MJ, Klazinga NS, Kringos DS. Statistics in Dutch policy debates on health and healthcare. Health Res Policy Syst. 2019;17(1):55.
5. Lavis JN, Ross SE, Hurley JE, Hohenadel JM, Stoddart GL, Woodward CA, et al. Examining the role of health services research in public policymaking. Milbank Q. 2002;80(1):125-54.
6. Kuruvilla S, Mays N, Walt G. Describing the impact of health services and policy research. J Health Serv Res Policy. 2007;12 Suppl 1:S1-23-31.
7. Bornmann L. Measuring the societal impact of research: research is less and less assessed on scientific impact alone--we should aim to quantify the increasingly important contributions of science to society. EMBO reports. 2012;13(8):673-6.
8. Greenhalgh T, Russell J. Evidence-based policymaking: a critique. Perspect Biol Med. 2009;52(2):304-18.
9. Smith R. Measuring the social impact of research. BMJ. 2001;323(7312):528.
10. Sumner P, Vivian-Griffiths S, Boivin J, Williams A, Bott L, Adams R, et al. Exaggerations and Caveats in Press Releases and Health-Related Science News. PLoS One. 2016;11(12):e0168217.
11. Schat J, Bossema FG, Numans ME, Smeets I, Burger P. Exaggerated health news. Nederlands Tijdschrift voor Geneeskunde. 2018;162(1).
12. Sumner P, Vivian-Griffiths S, Boivin J, Williams A, Venetis CA, Davies A, et al. The association between exaggeration in health related science news and academic press releases: retrospective observational study. BMJ : British Medical Journal. 2014;349.
13. Buhse S, Rahn AC, Bock M, Muhlhauser I. Causal interpretation of correlational studies - Analysis of medical news on the website of the official journal for German physicians. PLoS One. 2018;13(5):e0196833.
14. Haber N, Smith ER, Moscoe E, Andrews K, Audy R, Bell W, et al. Causal language and strength of inference in academic and media articles shared in social media (CLAIMS): A systematic review. PLoS ONE. 2018;13(5).
15. Alla K, Hall WD, Whiteford HA, Head BW, Meurk CS. How do we define the policy impact of public health research? A systematic review. Health Res Policy Sy. 2017;15(1):84-.
16. Dinsmore A, Allen L, Dolby K. Alternative Perspectives on Impact: The Potential of ALMs and Altmetrics to Inform Funders about Research Impact. PLOS Biology. 2014;12(11):e1002003.
17. Council for Medical Sciences. The societal impact of applied health research: Towards a quality assessment system. Amsterdam: Royal Netherlands Academy of Arts and Sciences; 2002.
18. Canadian Institutes of Health Research. Developing a CIHR Framework to Measure The Impact of Health Research. Ottawa; 2005.
19. Waddell C, Lomas J, Lavis JN, Abelson J, Shepherd CA, Bird-Gayson T. Joining the conversation: newspaper journalists' views on working with researchers. Healthcare policy = Politiques de sante. 2005;1(1):123-39.

20. Chapman S, Haynes A, Derrick G, Sturk H, Hall WD, St George A. Reaching “an audience that you would never dream of speaking to”: influential public health researchers’ views on the role of news media in influencing policy and public understanding. J Health Commun. 2014;19(2):260-73.
21. Gerrits RG, Jansen T, Mulyanto J, van den Berg MJ, Klazinga NS, Kringos DS. Occurrence and nature of questionable research practices in the reporting of messages and conclusions in international scientific Health Services Research publications: a structured assessment of publications authored by researchers in the Netherlands. BMJ Open. 2019;9(5):e027903.
22. McVay AB, Stamatakis KA, Jacobs JA, Tabak RG, Brownson RC. The role of researchers in disseminating evidence to public health practice settings: a cross-sectional study. Health Res Policy Syst. 2016;14(1):42.
23. Juttmann RE. Het domein van gezondheidszorgonderzoek. In: Plochg T, Juttmann RE, Klazinga NS, Mackenbach JP, editors. Handboek gezondheidszorgonderzoek. Houten: Bohn Stafleu van Loghum; 2007. p. 22-3.
24. Lohr KN, Steinwachs DM. Health services research: an evolving definition of the field. Health Serv Res. 2002;37(1):7-9.
25. Hsieh H-F, Shannon SE. Three Approaches to Qualitative Content Analysis. Qualitative Health Research. 2005;15(9):1277-88.
26. Gerrits RG, Kringos DS, van den Berg MJ, Klazinga NS. Improving interpretation of publically reported statistics on health and healthcare: the Figure Interpretation Assessment Tool (FIAT-Health). Health Res Policy Syst. 2018;16(1):20.
27. Adams RC, Challenger A, Bratton L, Boivin J, Bott L, Powell G, et al. Claims of causality in health news: A randomised trial. BMC Medicine. 2019;17(1).
28. Bossema FG, Burger P, Bratton L, Challenger A, Adams RC, Sumner P, et al. Expert quotes and exaggeration in health news: A retrospective quantitative content analysis. Wellcome Open Research. 2019;4.
29. Ioannidis JPA, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. The Lancet. 2014;383(9912):166-75.
30. Tijdink JK, Schipper K, Bouter LM, Maclaine Pont P, de Jonge J, Smulders YM. How do scientists perceive the current publication culture? A qualitative focus group interview study among Dutch biomedical researchers. BMJ Open. 2016;6(2):e008681.
31. Bouter LM. Commentary: Perverse Incentives or Rotten Apples? Accountability in Research. 2015;22(3):148-61.

7

Individual, institutional, and scientific environment factors associated with questionable research practices in the reporting of messages and conclusions in scientific Health Services Research publications

AUTHORS:

Reinie G. Gerrits, Joko Mulyanto, Joost D. Wammes , Michael J. van den Berg, Niek S. Klazinga, Dionne S. Kringos

Submitted for publication

ABSTRACT

BACKGROUND Health Services Research findings (HSR) reported in scientific publications may become part of the decision-making process on healthcare. This study aimed to explore associations between researcher's individual, institutional, and scientific environment factors and the occurrence of questionable research practices (QRPs) in the reporting of messages and conclusions in scientific HSR publications.

METHODS We employed a mixed-methods study design. We identified factors possibly contributing to QRPs in the reporting of messages and conclusions through a literature review, 14 semi-structured interviews with HSR institutional leaders, and 13 focus-groups amongst researchers. A survey corresponding with these factors was developed and shared with 172 authors of 116 scientific HSR publications produced by Dutch research institutes in 2016. We assessed the included publications for the occurrence of QRPs. An exploratory factor analysis was conducted to identify factors within individual, institutional, and environmental domains. Next, we conducted bivariate analyses using simple Poisson regression to explore factors' association with the number of QRPs in the assessed HSR publications. Factors related to QRPs with a p-value < .30 were included in four multivariate models tested through a multiple Poisson regression.

RESULTS In total, 78 (45%) participants completed the survey (51.3% first authors and 48.7% last authors). Twelve factors were included in the multivariate analyses. In all multivariate models, a higher score of "pressure to create societal impact" (Exp B = 1.28, 95% CI [1.11, 1.47]), was associated with higher number of QRPs. Higher scores on "specific training" (Exp B = 0.85, 95% CI [0.77-0.94]) and "co-author conflict of interest" (Exp B = 0.85, 95% CI [0.75-0.97]) factors were associated with a lower number of QRPs. Stratification between first and last authors indicated different factors were related to the occurrence of QRPs for these groups.

CONCLUSION Experienced pressure to create societal impact is associated with more QRPs in the reporting of messages and conclusions in HSR publications. Specific training in reporting messages and conclusions and awareness of co-author conflict of interests are related to fewer QRPs. Our results should stimulate awareness within the field of HSR internationally on opportunities to better support reporting in scientific HSR publications.

BACKGROUND

In 2009, it was estimated that 85% of research funding in biomedical sciences was avoidably wasted (1). In the biomedical sciences, evidence has been piling up on questionable research practices (QRPs) such as imbalanced research question selection, poor study design and execution, non-publication, and poor reporting (1). Over time, advancements have been made to address these QRPs, including scientific reporting (2). However, proper interpretation and reporting of messages and conclusions across different research methodologies in scientific publications requires more attention (3). Researchers can introduce various QRPs in the reporting of messages and conclusions in their scientific publications (e.g., generalizing findings to populations not included in the study, not reporting contradictory evidence, claiming an unjustified causal relationship, and inadequately justifying conclusions) (3-5). Moreover, although scientific reporting of biomedical studies is progressing (2), responsible scientific reporting requires greater awareness in the field of Health Services Research (HSR). HSR has a direct link to policy and practice, where stakeholders and funders may contribute considerably to the interpretation of results (6, 7). Additionally, HSR relies on mixed methodologies such as qualitative and mixed methods designs that may have less strict reporting requirements compared to quantitative designs such as randomized controlled trials (3).

Recent work has suggested that scientific HSR publications may include a median of six QRPs in the reporting of messages and conclusions (8). QRPs were primarily found in reported implications and recommendations for policy and practice, a lack of mention of contradictory evidence, and the conclusions of the scientific publication (8). The occurrence of these QRPs is concerning as messages and conclusions reported in scientific HSR literature are often transferred to policy makers, managers, and the general public. Further, these groups may learn about messages and conclusions directly from the scientific publication or through societal publications such as professional journals, factsheets, press releases, and reports (7, 9-12). Whether messages are disseminated by researchers, science communicators, or journalists, they may be accepted as established evidence and become part of the decision-making process on health and healthcare. Decisions on topics such as co-payments, adaption of protocols in hospitals, admitting medications to insurance packages, and tobacco regulation may thus be affected by inadequately reported messages and conclusions (9, 13, 14).

Scientific journals have taken the lead in implementing control measures to provide structure to the review process and improve responsible reporting

(9). These efforts have resulted in practices such as publication checklists (15), data sharing, open access (16) and public peer-review becoming increasingly common. Yet, these measures are primarily aimed at increasing transparency in reporting and thus may be insufficient in preventing QRPs in the reporting of messages and conclusions specifically. To strengthen the reporting of messages and conclusions, measures may need to be taken at multiple levels, including academic journals and research institutions themselves (18). Recently, Dutch academic and non-academic HSR institutions have begun to collaborate with the goal to increase responsible reporting of HSR findings. These efforts have been supported by the Netherlands Organization for Health Research and Development (ZonMw).

HSR institutions in the Netherlands have varying organizational policies in fostering responsible conduct of research, including responsible reporting. This variety in institutional culture and organisation offers the opportunity to learn from each other's reporting practices. Improving scientific publication of HSR requires an understanding of factors that influence authors in their writing, as well as those that impact the publication process itself (e.g., pressure and relationships with funders) (19-21). Research institutions may prevent the occurrence of QRPs by improving internal integrity and training researchers in scientific writing and communication (19-21). However, considering the specific characteristics of HSR, additional evidence is needed on how possible factors may relate to QRPs in messages and conclusions specifically (22).

Consequently, the aim of this study was to explore associations between individual, institutional, and scientific environment factors and the frequency of inadequacies in the reporting of messages and conclusions in scientific HSR publications.

METHODS

Design

We employed a mixed-methods study design. First, we identified factors possibly contributing to the occurrence of QRPs in the reporting of messages and conclusions in scientific HSR publications through a literature review, 14 semi-structured interviews with leaders of HSR groups or institutions in the Netherlands, and 13 focus-groups amongst junior health services researchers. Factors were clustered into three domains: individual, institutional and scientific environmental domains (9). Second, a survey corresponding to the identified factors was developed and shared with 172 first and last authors of a sample of 116 scientific HSR publications published in 2016 with an affiliation to Dutch HSR groups or institutions.

Setting

The study involved publications and participants from 13 HSR groups, departments, or institutions including both academic and non-academic institutions (hereafter referred to as "HSR institutions") in the Netherlands. These institutions agreed to participate in an effort to assure the overall quality of HSR publications in the Netherlands.

Conceptual framework on factors potentially associated with QRPs in the reporting of messages and conclusions in HSR

Factors potentially associated with QRPs in HSR were identified through an exploratory literature review, 14 semi-structured interviews with 19 leaders/representatives of the 13 participating institutions, and 13 focus-groups comprised of 57 junior/PhD researchers at participating HSR institutions. An initial overview of factors was created through the literature review. This overview was then discussed in the semi-structured interviews with the leaders/representatives of participating institutions. Within the focus-groups, an open conversation was held with participants to identify additional factors overlooked in the literature and interviews. Documented interview reports and transcripts were qualitatively analysed in MaxQDA resulting in the specification of factors potentially associated with QRPs in the reporting of messages and conclusions in scientific HSR publications. The applied methods for the development of these factors are described in more detail in supplementary material 1.

Identified factors were included in a theoretical framework consisting of three domains: individual, institutional, and scientific environments. Of note, factors within each domain could be influenced by those in other domains. The individual domain was comprised of factors bound to the individual researcher, including those associated with research experience and self-efficacy. The institutional domain included factors controlled by the institution that houses the researcher. These included institutional culture, facilities, interactions, and policies that may affect the writing and publication experience of the researcher. More concretely, an institution may have an (unofficial) policy to produce a certain number of publications per year. The scientific environmental domain included factors that manifest outside of the direct control of the institution, including those characterizing scientific culture and systems in general. Concrete examples of these factors were scientific journal policy (e.g., word length or use of reporting checklists) or collaboration between researchers and stakeholders outside the research institution.

Survey development

The survey was designed based on the framework described above. For each identified factor in the preliminary framework, one or more survey questions were developed. Questions were evaluated on their face validity by the co-authors and two project advisors, both senior health services researchers. For this study, we developed a new questionnaire, as no validated questionnaires were tailored to the field of HSR or scientific reporting. One existing question from the publication pressure questionnaire was included in our newly developed questionnaire (21). The questionnaire was developed in English (i.e., the primary working language of the study population).

A "think out loud" test was performed with two people from the target population. RG sat down with two researchers individually as they answered the survey questions and commented on their interpretation. The survey was designed in English and checked by a native speaker. After the final revision, the survey included 97 questions related to the factors within individual, institutional and scientific environment domains. Seven additional questions were included to assess personal and background characteristics. Answers to survey questions were provided on a Likert-scale (strongly disagree, disagree, neither agree nor disagree, agree, and strongly agree).

The survey is provided in supplementary material 2.

Survey study population

In a previous study we assessed QRPs in the reporting of messages and conclusions in 116 international peer-reviewed publications authored by researchers from the 13 participating institutions. QRPs were defined as "to report, either intentionally or unintentionally, conclusions or messages that may lead to incorrect inferences and do not accurately reflect the objectives, the methodology, or the results of the study." (8) For the assessment, we used a detailed assessment form including 35 possible QRPs in reporting messages and conclusions (e.g., "conclusions that do not adequately reflect the findings of the study", "limitations are not adequately justified"). This assessment form, along with corresponding methods and results, have been published elsewhere (6). For the current study, we conducted a survey amongst the 172 first and last authors of these publications.

First and last authors of the 116 scientific publications were included in our assessment.

We identified a total of 202 authors (116 first authors and 86 unique last authors) as the sample for our study. Contact information (i.e., e-mail addresses) was obtained through the participating institutions. These institutions were asked to encourage their researchers to participate in the survey, however, participation was voluntary and participants could stop at any time. Participants of the survey were informed of the goal of the study and data handling procedures in the invitation e-mail and at the start of the survey. We excluded 30 authors who's contact information was unknown, resulting in a final sample of 172 authors. The response rate was 45% (78 respondents).

Quantitative analysis

Dependent variable

The main dependent variable of this study was the number of QRPs in the reporting of messages and conclusions in HSR publications. The data were obtained from one of our previous studies (8).

Independent variables

The items (i.e., questions) included in the survey questionnaire derived from three major domains as described above (i.e., the individual, institutional, and scientific environment). An exploratory factor analysis was conducted to identify factors underlying items within each domain. The factors identified in the exploratory analyses were named as much as possible in alignment with the factors in our theoretical framework.

We used the factors identified from the factor analysis as independent variables in our analyses. The methods and results of the factor analysis are further described in supplementary material 3.

Considering the explorative nature of our study, no assumptions were made regarding the relative importance of factors within and between domains. Due to our explorative aim, we decided to include all factors resulting from the factor analyses.

Other characteristics

We collected several personal characteristics in the survey (i.e., age, working experience as scientific researcher, academic background, academic position, number of publications co-authored, and journal's impact factor). We described sample characteristics based on these variables.

Statistical analysis

The basic characteristics of study samples were described based on the measurement scale of the variables. Categorical variables (nominal / ordinal) were presented as frequency and percentage, whereas numerical variables (interval/ratio) were presented using mean and standard deviation.

We conducted bivariate analysis using simple Poisson regression. Poisson regression was chosen considering the nature of the outcome (number of QRPs) as count data with a relatively small mean value. This analysis specifically assessed the association between each factor score and the number of QRPs in HSR publications. The analysis was also intended to reduce the number of factors which were included in the multivariate model.

Following the bivariate analysis, we applied multiple Poisson regression to further assess the association between the factor domains and the number of QRPs in HSR publications. For the purpose of model development, we provided four models in our multivariate analysis to ensure the stability of our results. The first two models were crude models (unadjusted) and included 12 factors from the bivariate analyses that exhibited a significant association with QRPs (*ps* < .3). The last two models included number of years of work experience as scientific researcher, as well as the journal's impact factor, to examine the influence of these variables' influence on the quality of reporting. For easier interpretation, we provided the coefficient of each explanatory variable (B), exponential form of the coefficient (Exp B) and the 95% confidence interval (95% CI). The goodness of fit of all models was checked using the chi-square goodness of fit test as part of the Poisson regression procedure, with results suggesting all models demonstrated good fit.

Considering that the factors we used as independent variables may be interrelated, we checked for collinearity in our regression model. Results from the correlation matrix in the exploratory factor analysis procedure showed 35 of 171 pairs between scales (20%) were significantly but not strongly correlated (*rs* < 0.3). Hence, these findings suggested no multi-collinearity issues in our analysis.

Because first and last authors have different roles in the writing of scientific publications, we provided additional stratified analysis between first and last authors to further explore the nature of the association between these factors and the number QRPs. All analyses were conducted using IBM SPSS version 25.

Ethics approval

A waiver for ethical approval was obtained for this study from the medical ethics review committee at Amsterdam UMC. To avoid negative consequences for participants, each participant and publication was assigned a unique identification number. Extracted data were entered in SPSS using this number to separate author information from the study data.

RESULTS

Of the survey participants, 51.3% were a first author and 48.7% were a last author. PhD students (25.6%) and professors (29.5%) were the most frequent academic positions in our sample. First authors were predominantly PhD students (50.0%), whereas the last authors were predominately professors (57.9%). Both first (40%) and last (28.9%) authors primarily had an academic background in the social sciences (40%). Last authors were older, had longer working experience as a scientific researcher, and reported a larger number of publications co-authored as compared to first authors. The journal impact factor of the publications was similar between last and first authors. The number of QRPs per publication was slightly higher for the last authors than first authors. The basic characteristics of the study sample are provided in Table 1. There are 12 publications corresponding to both first and last author, 28 publications correspond only to a first author, and 26 publications correspond only to a last author.

Bivariate analyses

Table 2 depicts findings from bivariate analyses examining the relationship between each factor from the individual, institutional, and scientific environment domain and the number of QRPs. Of the five factors in the individual domain, “pressure to create societal impact” (Exp B = 1.34, 95% CI [1.18, 1.51]) and “self-efficacy” (Exp B = 0.84, 95% CI [0.72, 0.98]) exhibited significant associations with the number of QRPs. For institutional factors, only “specific training in reporting messages and conclusions” (Exp B = 0.85, 95% CI [0.77, 0.93]) exhibited a significant association with the number of QRPs. “Stakeholder influence (Exp B =1.16, 95% CI [1.06, 1.27]) was the only factor from the scientific environmental domain that exhibited a significant association with the number of QRPs.

Table 1 | Basic characteristics of survey respondents.

	Overall		First author		Last author	
	n	%	n	%	n	%
Author status						
First author	40	51.3	-	-	-	-
Last author	38	48.7	-	-	-	-
Research position (January 2016)						
PhD student	20	25.6	20	50.0	0	0.0
Post-doctoral researcher	10	12.8	10	25.0	0	0.0
Senior researcher	13	16.7	5	12.5	8	21.1
Assistant professor	6	7.7	2	5.0	4	10.5
Associate professor	5	6.4	1	2.5	4	10.5
Professor	23	29.5	1	2.5	22	57.9
Other	1	3.1	1	2.5	0	0.0
Academic background						
Social sciences	27	34.6	16	40.0	11	28.9
Epidemiology	21	26.9	10	25.0	11	28.9
(Health) economics	4	5.1	2	5.0	2	5.3
Other	26	33.3	12	30.0	14	36.8
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Age of participant	45.21	11.67	37.90	10.08	52.92	7.60
Working experience as scientific researcher	15.83	9.78	9.44	6.99	22.55	7.63
Average number of publications co-authored per year	4.88	1.95	3.43	1.47	6.42	0.98
Journal Impact Factor of publication	2.07	1.74	2.08	1.50	2.06	1.99
Number of QRPs in publication	6.04	3.46	5.83	3.50	6.26	3.45

Multivariate analyses

Results of multivariate analyses are presented in Table 3. Of the four models in our analysis, three factors i.e., "pressure to create societal impact", "specific training", and "co-author conflict of interest" consistently exhibited a significant association with the number of QRPs.

In the fully-adjusted model (i.e., model 3), a one-point increase on the "pressure to create societal impact" item was associated with a 28% increase in the number QRPs in an HSR publication (Exp B = 1.28, 95% CI [1.11, 1.47]). Conversely, a one-point increase on the "specific training in reporting messages and conclusions" was associated with a 15% decrease in the number QRPs of an HSR publication (Exp B = 0.85, 95% CI [0.7, -0.94]). A one-point increase on the "co-author conflict of interest" item was associated with a 15% decrease in the number of QRPs in an HSR publication (Exp B = 0.85, 95% CI [0.75, 0.97]).

Table 2 | Results of bivariate Poisson regression analysis examining individual, institutional, and scientific environment domain factors' associations with number of QRPs.

Domain	Factors	B	SE	p-value	Exp(B)	95% CI
Individual	Ambition in science	0.075	0.090	0.404	1.08	0.90-1.28
	Self-efficacy	-0.171	0.079	0.031	0.84	0.72-0.98
	Perception of received training	-0.008	0.081	0.917	0.99	0.85-1.16
	Confidence in writing	-0.111	0.083	0.180	0.89	0.76-1.05
	Pressure to create societal impact	0.289	0.064	0.000	1.34	1.18-1.51
	Perception of contribution to science.	0.102	0.069	0.139	1.11	0.97-1.27
Institution	Specific training in reporting messages and conclusions	-0.169	0.048	0.000	0.85	0.77-0.93
	Competitiveness	0.089	0.053	0.091	1.09	0.98-1.21
	Data storage	-0.045	0.054	0.404	0.96	0.86-1.06
	Feedback culture at institute	0.022	0.051	0.666	1.02	0.93-1.13
	Social support	-0.034	0.084	0.683	0.97	0.82-1.14
	Media policy	0.018	0.059	0.757	1.02	0.91-1.14
	Influence of funders	-0.068	0.065	0.296	0.93	0.82-1.06
Environment	Creating exciting conclusion	0.019	0.070	0.783	1.02	0.89-1.17
	Media contact	-0.018	0.061	0.769	0.98	0.87-1.11
	Pressure from scientific culture	0.048	0.071	0.495	1.05	0.91-1.21
	Suspicions of co-workers	-0.100	0.071	0.160	0.91	0.79-1.04
	Journal practice	-0.089	0.067	0.186	0.92	0.80-1.04
	Stakeholder influence	0.146	0.047	0.002	1.16	1.06-1.27
	Co-author conflict of interest	-0.100	0.057	0.082	0.91	0.81-1.01
	Conflict between co-authors	-0.077	0.045	0.087	0.93	0.85-1.01

Stratified analyses between first and last authors

Results from stratified analyses between first and last authors, along with results of multivariate analyses using the fully adjusted model, are included in Table 4. A complete description of our stratified analysis with all applied models can be found in the supplementary material 4.

For first authors, findings indicated "specific training" reporting messages and conclusions (Exp B = 0.84, 95% CI [0.72, 0.98]) was associated with fewer QRPs. "Feedback culture" at their research institute (Exp B = 1.24, 95% CI [1.05, 1.47]) and "pressure to create societal impact" (Exp B = 1.24, 95% CI [1.02, 1.51]) contribute to a higher number of QRPs. For last authors no significant relationship was identified between factors and QRPs.

Table 3 | Multivariate analysis between factors from individual, institutional, and scientific environment domains with number of QRPs using Poisson regression.

	Model 1^{a}			Model 2^{b}			Model 3^{a}			Model 4^{b}		
	B	Exp(B)	95% CI	B	Exp(B)	95% CI	B	Exp(B)	95% CI	B	Exp(B)	95% CI
Intercept	1.397	4.04	1.24-13.10	1.501	4.48	2.07-9.72	1.633	5.11	1.50-17.45	1.634	5.12	2.27-11.54
Journal impact factor	-	-	-	-	-	-	0.051	0.95	0.89-1.01	-0.053	0.95	0.89-1.01
Working duration	-	-	-	-	-	-	-0.001	0.99	0.98-1.01	-0.000	1.00	0.99-1.01
Individual												
Self efficacy	-0.058	0.94	0.78-1.13	-0.130	0.98	0.75-1.03	-0.055	0.95	0.79-1.14	-0.129	0.88	0.75-1.03
Confidence in writing	-0.029	0.97	0.81-1.17	-	-	-	-0.018	0.95	0.78-1.15	-	-	-
Pressure to create societal impact	0.268	1.31	1.14-1.50	0.226	1.25	1.10-1.42	0.245	1.28	1.11-1.47	0.199	1.22	1.07-1.39
Perception of contribution to science.	0.100	1.11	0.94-1.30	-	-	-	0.084	1.09	0.93-1.28	-	-	-
Institution												
Specific training	-0.171	0.84	0.76-0.93	-0.155	0.86	0.78-0.94	-0.165	0.85	0.77-0.94	-0.147	0.86	0.78-0.95
Competitiveness	0.057	1.06	0.95-1.18	-	-	-	0.052	1.05	0.94-1.18	-	-	-
Influence of funders	-0.078	0.93	0.79-1.08	-	-	-	-0.075	0.93	0.80-1.08	-	-	-
Environment												
Suspicion of co-workers	-0.002	0.99	0.84-1.18	-	-	-	0.002	1.01	0.85-1.20	-	-	-
Journal practice	-0.027	0.97	0.84-1.13	-	-	-	-0.022	0.98	0.84-1.13	-	-	-
Stakeholder influence	0.050	1.05	0.95-1.17	0.091	1.10	0.99-1.21	0.065	1.07	0.96-1.19	0.103	1.11	1.00-1.22
Co-author conflict of interest	-0.151	0.86	0.76-0.98	-	-	-	-0.157	0.85	0.75-0.97	-	-	-
Conflict between co-authors	-0.016	0.98	0.89-1.10	-	-	-	-0.012	0.98	0.89-1.09	-	-	-

^{a} *Included domains with p-value <0.30 in bivariate analysis;* ^{b}*Included domains with p-value <0.05 in bivariate analysis;*

Table 4 | Comparison of factors associated with the number of QRPs in reporting of messages and conclusions in HSR publication between first and last authors

Factors[a]	First author			Last author		
	B	Exp(B)	95% CI	B	Exp(B)	95% CI
Intercept	0.248	1.28	0.18-0.93	2.212	9.33	1.47-56.60
Journal impact factor	-0.114	0.89	0.80-1.01	-0.066	0.94	0.86-1.02
Working duration	0.004	1.01	0.98-1.03	0.005	1.01	0.98-1.02
Individual						
Ambition in science	0.306	1.36	0.99-1.85	-	-	-
Self-efficacy	-	-	-	-0.198	0.82	0.59-1.14
Perception of received training	-0.139	0.87	0.68-1.14	0.253	1.29	0.94-1.75
Confidence in writing	-	-	-	-0.360	0.70	0.48-1.01
Pressure to create societal impact.	0.216	1.24	1.02-1.51	0.175	1.19	0.90-1.58
Perception of contribution to science.	0.192	1.21	0.94-1.55	-	-	-
Institution						
Specific training	-0.170	0.84	0.72-0.98	-0.058	0.94	0.77-1.15
Competitiveness	-	-	-	0.099	1.10	0.88-1.39
Data storage	-0.112	0.89	0.73-1.10	-	-	-
Feedback culture at institute	0.218	1.24	1.05-1.47	-0.094	0.91	0.75-1.10
Influence of funders	-0.173	0.84	0.70-1.02	-	-	-
Environment						
Creating exciting conclusion	0.060	1.06	0.79-1.44	-0.052	0.95	0.73-1.24
Suspicions of co-workers	-	-	-	0.047	1.05	0.76-1.44
Journal practice	-0.075	0.93	0.77-1.12	-	-	-
Stakeholder influence	0.052	1.05	0.90-1.24	-0.012	0.98	0.77-1.27
Co-author conflict of interest	-	-	-	-0.083	0.92	0.75-1.12
Conflict between co-authors	-	-	-	-0.118	0.89	0.74-1.06

[a] *The multivariate models included different factors (independent variables) between first author and last author analysis resulting the bivariate analysis. For a complete description please refer to the supplementary material 4.*

DISCUSSION

The aim of this study was to explore the possible association between individual, institutional, and scientific environment factors and inadequacies in the reporting of messages and conclusions in scientific HSR publications.

We identified three factors independently associated with QRPs in the reporting of messages and conclusions in scientific HSR publications i.e., “pressure to create societal impact”. “specific training reporting messages and conclusions" and "co-author conflict of interest". Stratification between first and last author indicated different factors related to the occurrence of QRPs.

Limitations

The main strength of our approach was our mixed methods design. By constructing a framework from the experiences of a sample of health services researchers, we could tailor the survey to our study participants. Moreover, most research on research integrity is derived from self-report. Our assessment of QRPs provides a more impartial approach.

Considering the large turnover of research staff and PhD students at each institution, a response of 45% may be optimal. The average number of QRPs is similar to that of all assessed HSR publications. Nevertheless, non-respondents might have rejected participation because of time pressures or a lack of communication with the HSR community. The relatively small sample size in the current study also presents as a limitation and necessitates replication in larger and more diverse samples. Further, due to our explorative aim, we decided to include factors with a lower threshold of reliability. In follow-up research we recommend to improve the factors' reliability.

We acknowledge the publications we analysed are nested in the thirteen participating Dutch institutions, which may influence the associations between institutional factors and QRPs. A multilevel analysis including the institution in which the publication is nested would be the ideal option to address this issue. However, such an analysis would likely have required a larger dataset to provide a robust estimation of effects, particularly given the fact that publications are typically written by authors from multiple institutions. Considering the relatively small sample size and the explorative nature of our study, a single-level regression was a more appropriate choice. Further study with a larger data set and clearly distinguished institutions will allow for a more sophisticated analysis technique to confirm findings from our study.

The assessed publications were published in 2016. Researchers might not have had a very vivid memory of their working experience two years prior. This would have made our connections between publication and person somewhat less reliable. Nevertheless, we do not expect institutional or scientific factors to have changed significantly in the course of two years. Risk for recall bias was likely minimal.

The researchers and studies included in this study all originated from Dutch research institutions. Although institutional structures and individual experiences of research culture will be different across countries, HSR researchers and institutions often deal with similar challenges as those encountered in the Netherlands, including publication pressures and creating societal impact. Aspects of the results from the current study are thus likely to provide a helpful guide for HSR institutions internationally.

Interpretation

Our results indicated three factors are independently associated with QRPs in the reporting of messages and conclusions in HSR literature. The other factors in the assessed framework, however, are not irrelevant. All included factors may relate to multiple aspects of the publication process and are worth addressing in future studies on QRPs. Our study was explorative, and we therefore recommend further empirical research on the resulting factors.

The association between a higher number of QRPs and the factor "pressure to create societal impact" facilitates important insights on the current research culture of HSR. HSR is often intended for practical intervention (23). To improve the connection between HSR and policy and practice across the entire field of HSR, researchers are stimulated to spread their findings via societal publications to policy makers, professionals and the public (24). Researchers may anticipate their societal impact when writing their scientific publications. Hence, they may be likely to unconsciously adapt their language and writing to present concrete and actionable conclusions suited to attract the attention of the media or the professional community (25). It is generally assumed that pressure to create societal impact pushes authors to overstate conclusions in press releases or other societal publications. However, the current findings suggest a possible effect on scientific reporting as well. Currently, researchers may not have the means to responsibly create societal impact or have difficulty aligning their scientific messages with societal messages. Not all researchers are equally equipped for this task. Future studies addressing scientific reporting should take into account the association between the perception of research impact and reporting in scientific publications.

"Specific training in reporting messages and conclusions" was associated with a lower number of QRPs. The positive association between training and the improvement of writing skills has been identified in previous studies (26, 27). Because the participants self-reported on their level of specific training, our findings highlight that some courses offered by HSR institutions in the Netherlands may provide researchers with helpful tools to improve their writing. Moreover, researchers may be capable in recognizing they need more specific training. Institutions should assure that those who need specific training in reporting messages and conclusions will be able to obtain it.

“Co-author conflict of interest” was associated with a lower number of reporting adequacies. This finding contradicts the assumption that research quality generally decreases when a conflict of interest arises. A possible explanation may be that awareness of a conflict of interest by co-authors may have stimulated a more nuanced or careful interpretation of the research findings. Policies in place at HSR institutions could assure that those conflicts of interests are positively mitigated and result in more attention to research conduct (28).

One method used by research institutions to force a stimulating debate is to introduce structured peer-feedback (29). Although some institutions in the Netherlands have invested in structuring feedback support for their researchers, feedback culture was not associated with a lower number of QRPs in the current study. Surprisingly, the analyses differentiating between first and last authors indicated that feedback culture may contribute to more QRPs for first authors. Feedback structures differ for each institution, and some might not be aimed sufficiently at the interpretation and reporting of messages and conclusions. It could thus be worthwhile to investigate how feedback structures can better support authors and what type of feedback culture would specifically create a stimulating debate. The assessment form developed for assessing QRPs in scientific publications might guide structured feedback on reported messages and conclusions specifically (6).

Our analyses further indicate that factors contributing to QRPs may be different for first and last authors. First authors may contribute to more QRPs when they experience more pressure to create societal impact and a positive feedback culture. They may contribute to fewer QRPs when they receive more specific training in reporting messages and conclusions. Additional research on the unique roles of first and last authors in the prevention of QRPs when reporting messages and conclusions in scientific HSR publications is recommended.

Implications and recommendations for policy and practice

Our study identified factors that are best addressed through changes by research institutions. Results should stimulate awareness within the HSR community internationally. In support of a more responsible translation of findings to policy and practice, they should address the identified factors to contribute to better reporting in scientific HSR publications. We recommend the development of institutional interventions to encourage responsible reporting of messages and conclusions in HSR. Specialized writing courses and workshops may increase writing confidence(27). Specific training already in place at research institutions on writing discussions and conclusions should be extended to all health services researchers who do not have access. HSR institutions should further prioritize providing a positive feedback culture by stimulating debate and making conflicting interests explicit. They should moreover, introduce systematic changes such as organizing peer-review with engaging discussions, and providing sufficient time and support in balancing scientific reporting and creating societal impact. Across the HSR field, institutions are already taking actions to assure responsible research practices, thus we recommend them to strengthen the coherence of their efforts, also by collaboration.

CONCLUSION

Experienced pressure to create societal impact is associated with a higher number of QRPs in the reporting of messages and conclusions in HSR publications. Specific training in reporting messages and conclusions, and awareness of co-author conflict of interests are related to fewer QRPs in HSR publications. This study was exploratory and we therefore recommend further research on the identified factors. Our results should stimulate awareness within the field of HSR internationally on opportunities better support reporting in scientific HSR publications, and thus a more responsible translation of findings to policy and practice.

REFERENCES

1. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, et al. Reducing waste from incomplete or unusable reports of biomedical research. The Lancet. 2014;383(9913):267-76.
2. Evolution of poor reporting and inadequate methods over time in 20 920 randomised controlled trials included in Cochrane reviews: research on research study. BMJ. 2017;358:j3806.
3. Boutron I, Ravaud P. Misrepresentation and distortion of research in biomedical literature. Proc Natl Acad Sci U S A. 2018;115(11):2613-9.
4. Fletcher RH, Black B. Spin in Scientific Writing: Scientific Mischief and Legal Jeopardy Defining Scientific Misconduct. Medicine and Law. 2007;26:511-26.
5. Horton R. The rhetoric of research. BMJ. 1995;310(6985):985.
6. Reijmerink W, Robben P, Ruwaard D, Vermeulen H. Naar gezond zorgonderzoek: interactieve leerervaringen uit het ZonMw-programma Gezondheidszorgonderzoek. Tijdschrift voor gezondheidswetenschappen. 2014;92(8):319-22.
7. Lavis JN, Ross SE, Hurley JE, Hohenadel JM, Stoddart GL, Woodward CA, et al. Examining the role of health services research in public policymaking. Milbank Q. 2002;80(1):125-54.
8. Gerrits RG, Jansen T, Mulyanto J, van den Berg MJ, Klazinga NS, Kringos DS. Occurrence and nature of questionable research practices in the reporting of messages and conclusions in international scientific Health Services Research publications: a structured assessment of publications authored by researchers in the Netherlands. BMJ Open. 2019;9(5):e027903.
9. Burstein P. The impact of public opinion on public policy: A review and an agenda. Polit Res Quart. 2003;56(1):29-40.
10. Hanney SR, Gonzalez-Block MA, Buxton MJ, Kogan M. The utilization of health research in policy-making: concepts, examples and methods of assessment. Health Res Policy Syst. 2003;1.
11. Kuruvilla S, Mays N, Walt G. Describing the impact of health services and policy research. J Health Serv Res Policy. 2007;12 Suppl 1:S1-23-31.
12. Gerrits RG, van den Berg MJ, Klazinga NS, Kringos DS. Statistics in Dutch policy debates on health and healthcare. Health Res Policy Syst. 2019;17(1):55.
13. Bouter LM. Commentary: Perverse Incentives or Rotten Apples? Accountability in Research. 2015;22(3):148-61.
14. Caulfield T. The Commercialisation of Medical and Scientific Reporting. PLoS Med. 2005;1(3):e38.
15. Altman DG, Simera I, Hoey J, Moher D, Schulz K. EQUATOR: reporting guidelines for health research. Lancet. 2008;371.
16. Moher D, Glasziou P, Chalmers I, Nasser M, Bossuyt PMM, Korevaar DA, et al. Increasing value and reducing waste in biomedical research: who's listening? The Lancet. 2016;387(10027):1573-86.
17. Opening up peer review. Nature. 2018;560(7720):527.
18. Schrag NJ, Purdy GM. Step up for quality research. Science. 2017;357(6351):531.
19. Haven TL, Bouter LM, Smulders YM, Tijdink JK. Perceived publication pressure in Amsterdam: Survey of all disciplinary fields and academic ranks. PLoS ONE. 2019;14(6).
20. Crain AL, Martinson BC, Thrush CR. Relationships between the Survey of Organizational Research Climate (SORC) and self-reported research practices. Sci Eng Ethics. 2013;19(3):835-50.

21. Tijdink JK, Verbeke R, Smulders YM. Publication pressure and scientific misconduct in medical scientists. J Empir Res Hum Res Ethics. 2014;9(5):64-71.
22. O'Cathain A, Murphy E, Nicholl J. The quality of mixed methods studies in health services research. J Health Serv Res Policy. 2008;13(2):92-8.
23. Bentley P, Kyvik S. Academic staff and public communication: a survey of popular science publishing across 13 countries. Public Understanding of Science. 2010;20(1):48-63.
24. Council for Medical Sciences. The societal impact of applied health research: Towards a quality assessment system. Amsterdam: Royal Netherlands Academy of Arts and Sciences; 2002.
25. Weingart P. Science and the media. Res Policy. 1998;27(8):869-79.
26. Rickard CM, McGrail MR, Jones R, O'Meara P, Robinson A, Burley M, et al. Supporting academic publication: Evaluation of a writing course combined with writers' support group. Nurse Education Today. 2009;29(5):516-21.
27. Kramer B, Libhaber E. Writing for publication: Institutional support provides an enabling environment. BMC Medical Education. 2016;16(1).
28. Besley JC, McCright AM, Zahry NR, Elliott KC, Kaminski NE, Martin JD. Perceived conflict of interest in health science partnerships. PloS one. 2017;12(4):e0175643-e.
29. Fanelli D, Costas R, Lariviere V. Misconduct Policies, Academic Culture and Career Stage, Not Gender or Pressures to Publish, Affect Scientific Integrity. PLoS One. 2015;10(6):e0127556.

SUPPLEMENTARY MATERIAL 1

This document describes the qualitative methods on the development of a framework of factors contributing to reporting inadequacies in messages and conclusions in Health Services Research (HSR).

Factors potentially associated with reporting inadequacies in HSR were identified through an explorative review of scientific literature, 14 semi-structured interviews with leaders/representatives (n=19) of 13 participating HSR institutions and 13 focus groups with junior/PhD researchers (n=57) at the participating HSR institutions.

Explorative literature review

First, an explorative literature review was conducted searching for factors related to the conduct of questionable research practices in general. Initial search terms included in different order and combination: 'questionable research practices', 'factors', 'questionable conclusions', 'misconduct' and 'spin'. The majority of publications were included through snowball sampling. Publications were included if they mentioned possible factors that might impact the conduct of questionable research practices, or provided definitions of the above key terms.

After identifying a body of literature, we screened all publications for factors possibly related to questionable research practices, and in particular might relate to the reporting of messages and conclusions. A broad selection of factors was included in an initial framework. This framework was used as a basis for semi-structured interviews with the leaders and representatives of the thirteen participating institutions. The publications identified are listed at the end of this supplement.

Semi-structured interviews

Second, nineteen leaders/representatives of the thirteen participating HSR institutions were interviewed during fourteen semi-structured interviews. Two interviews took place at one institution as it was represented by two departments. Three interviews were conducted with both the institute leader and a second representative. One of the interviews included three representatives of an institution. The aim of the interviews was to discuss our draft of Questionable Research Practices (QRPs), referred to in this paper as 'reporting inadequacies', and identify additional measurable reporting inadequacies in the reporting of messages and conclusions in HSR, explore potential causes of reporting inadequacies in messages and conclusions, and to discuss experiences of the institute leaders with these reporting inadequacies. A semi-structured interview guide was developed by the project team, that is published as an attachment to Gerrits et al. (2019) in BMJ open[1] We presented the interviewees with the preliminary framework of factors identified in the literature. The draft framework was iteratively adjusted, i.e. after each interview we drafted a new version including the findings of the previous interviews, which we then presented during each following interview.

Interviewees were approached through e-mail to schedule an appointment. Two researchers conducted the interviews of which thirteen took place at the participating institutions and one interview took place in a public space. During the first interview, both interviewers were present to align their interviewing approach. The remainder of the interviews were equally divided between the interviewers. The interviews lasted one hour. In concordance with ethical guidelines, the goal of the interview was explained at the start of the interview and permission to audio-record the interview was obtained.

With the support of the recordings, a report was written and shared with the interviewees for validation. From these validated reports and the updated list of reporting inadequacies, a final list of reporting inadequacies was drafted.

---

[1] Gerrits RG, Jansen T, Mulyanto J, van den Berg MJ, Klazinga NS, Kringos DS. Occurrence and nature of questionable research practices in the reporting of messages and conclusions in international scientific Health Services Research publications: a structured assessment of publications authored by researchers in the Netherlands. BMJ Open. 2019;9(5):e027903.

Focus groups

Third, we applied a focus group approach to explore possible factors related to the occurrence of reporting inadequacies in HSR.

Recruitment of focus group participants

The thirteen participating institutions were asked to support the organization of these focus groups. One institution declined as they were engaged in education during the research period. The intended duration of the focus groups was one and a half hour. The intended size of the focus groups was six to ten. Focus group participants were recruited through a contact person at the respective institutions. These contact persons were provided with a full description of the aims and methods of the focus groups. When potential participants were identified, a date for the focus group was determined. We emphasized that participation was completely voluntary, and no requirement by the participating institutions.

In total,13 focus groups at 12 institutions (in one institute two departments participated) were conducted, including 57 participants. The average number of participants was 4-5.

Setting

All focus group discussions took place at the institute locations to ease traveling expenses and time of the participants. Focus groups were conducted between April 2018 and October 2018.

During 11 focus group two researchers were present, one acting as facilitator and one as note taker. The facilitator ensured that the session progressed fluently, while the note taker took notes of the content and made sure that all the topics are covered. During the two last focus groups, one researcher was present, who both facilitated and took notes of the conversation. Participants were fully informed on the goal of the study. They were informed through e-mail. At the start of each focus group, the goal of the study was explained and it was clearly expressed participation was voluntary. Consent to participate by all participants was confirmed at the start of each focus group.

With permission of the focus group participants all focus groups were audiotaped. Ten recordings were transcribed ad verbatim. For all thirteen focus groups, a report was written. All participants were sent their respective reports for validation by e-mail.

The focus groups were guided by a focus group guide addressing the experience of researchers on factors that might influence reporting inadequacies, experienced good practices and discovering new good practices in the field of HSR. The transcriptions were analysed through content analyses. The results from this analysis was confirmed with the reports of the remaining three focus groups.

Analyses

The transcripts were analysed by inductive coding. The following steps were used to systematically analyse the transcribed data; (1) raw data files were prepared in the analyses program MaxQDA, (2) the full transcriptions were read closely, (3) themes were created, (4) overlapping codes and themes were connected (5) the themes were iteratively refined. Throughout the whole coding process, emerging themes were regularly compared and discussed by the two researchers and the research group.

First, JW and RG independently analysed two transcripts. They then compared their codes and agreed upon an initial coding scheme. JW then continued with the analyses of the eight remaining transcripts. RG regularly checked the coding performed by JW. After the analyses of 10 focus group transcripts, saturation was reached. RG validated the results from the analyses with the focus group reports from the last three focus groups.

Factors resulting from the focus group analyses were added to the framework resulting from the semi-structured interviews with leaders and representatives of the institutions.

Ethics

A waiver for ethical approval was obtained from the METC at the Academic Medical Center.

Framework

Identified factors were included in a theoretical framework existing of three domains: the individual, institutional, and scientific environment domain. The individual domain concerns characteristics that are bound to the individual researcher. These may concern characteristics such as research experience and self-efficacy. The institutional domain includes factors that are controlled by the institution that houses the researcher. These include institutional culture, facilities, interactions and policies that may affect the writing and publication experience of the researcher. For example, an institution may have an (unofficial) policy to stimulate a certain number of publications per year. The scientific environmental domain includes those factors that take place outside of the control of the institution, and belong to the scientific culture or system a researcher is part of. These are the factors that institutions have no direct control over. For example, factors are scientific journal policy including word length or use of reporting checklists, or collaboration with researchers and stakeholders outside the researching institution. Factors within each domain may be influenced by factors in other domains.

Framework and included factors

1 Scientific environment domain	2 Institutional domain	3 Individual domain
Funding • Funding rewards innovation & novelty • Demands of the funder **Valorisation of research outcomes** • Revenue model • Public media • Media pressure **Policies & practices scientific society** • Competition for research positions • Journal policies & practices • Peer review process • Pressure to publish 'exciting' articles **Collaborating partners** • Conflicts of interest **Research beneficiaries / stakeholders** • Usefulness: study designed without proper consideration of the value for e.g. patients	**Structural conditions/ resources** • Education • Reward system / incentives • Presence and adherence to a Research code • Recruitment & selection researchers • Presence of formal quality policy • Transparency study materials/data **Social conditions** • Opportunities for peer-discussion • Presence of colloquia for article discussion • Review of pre-publication findings • Competitiveness **Role of supervisors** • Task perception • Workload • Social skills (in supervision, and collaboration) **Cultural conditions** • Ideology institute • Functioning of the ICT infrastructure • Open organisation culture	**Motivation** • Promotion • Respect from peers • Focus on short-term success • Ideology **Capabilities** • Research training in HSR • Writing skills • Training research integrity • Social skills • Self-efficacy (to stand up to pressure) **Working conditions** • Workload • Work pressure **Perceptions** • Self-perception • Perception of others **Personality traits** • Narcissism

Publications identified in the exploratory review

Asman, O., Melnikov, S., Barnoy, S., & Tabak, N. (2017). Experiences , behaviors , and perceptions of registered nurses regarding research ethics and misconduct. *SAGE*, 1–11.

Bohannon, J. (2013) Who's Afraid of Peer Review? Science magazine 342 (6154): 60-65.

Bosch, X. (2014). Improving biomedical journals ' ethical policies: the case of research misconduct. *Journal of Medical Ethics*, *40*, 644–646. https://doi.org/10.1136/medethics-2013-10182

Bouter, L. M. (2015). Commentary: Perverse incentives or rotten apples?. *Accountability in research*, *22*(3), 148-161.

Braakhekke, M., Scholten, I., Mol, F., Limpens, J., Mol, B. W., & Veen, F. Van Der. (2017). Selective outcome reporting and sponsorship in randomized controlled trials in IVF and ICSI. *Human Reproduction*, *32*(10), 2117–2122. *https://doi.org/10.1093/humrep/dex273*

Chalmers, I., & Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *Obstetrics & Gynecology*, *114*(6), 1341-1345.

Crain, A. L., Martinson, B. C., & Thrush, C. R. (2013). Relationships between the Survey of Organizational Research Climate (SORC) and self-reported research practices. *Science and engineering ethics*, *19*(3), 835-850.

English, T., Antes, A. L., Baldwin, K. A., & Dubois, J. M. (2017). *Development and Preliminary Validation of a New Measure of Values in Scientific Work*. *Science and Engineering Ethics*. Springer Netherlands. https://doi.org/10.1007/s11948-017-9896-0

Fanelli, D., Costas, R., & Ioannidis, J. P. A. (2017). Meta-assessment of bias in science. *PNAS*, *114*(14). https://doi.org/10.1073/pnas.1618569114

Glasziou, P., Altman, D. G., Bossuyt, P., Boutron, I., Clarke, M., Julious, S., ... & Wager, E. (2014). Reducing waste from incomplete or unusable reports of biomedical research. *The Lancet*, *383*(9913), 267-276.

Godecharle, S., Nemery, B., & Dierickx, K. (2014). Heterogeneity in European research integrity guidance: relying on values or norms?. *Journal of Empirical Research on Human Research Ethics*, *9*(3), 79-90.

Haddaway, N. R., Woodcock, P., Macura, B., & Collins, A. (2015). Making literature reviews more reliable through application of lessons from systematic reviews. *Conservation Biology*, *29*(6), 1596–1605. JOUR. https://doi.org/10.1111/cobi.12541

Horrobin, D. F. (1999). Beyond conflict of interest. Non-financial conflicts of interest are more serious than financial conflicts. *The BMJ*, *318*(7181), 466-466.

Horton, R. (1995). The rhetoric of research. *BMJ: British Medical Journal*, *310*(6985), 985.

Ioannidis, J. P., & Khoury, M. J. (2014). Assessing value in biomedical research: the PQRST of appraisal and reward. *Jama*, *312*(5), 483-484.

Ioannidis, J., Tatsioni, A., & Karassa, F. B. (2010). Who is afraid of reviewers' comments? Or, why anything can be published and anything can be cited. *European journal of clinical investigation*, *40*(4), 285-287.

Jefferson, T., Di Pietrantonj, C., Debalini, M. G., Rivetti, A., & Demicheli, V. (2009). Relation of study quality, concordance, take home message, funding, and impact in studies of influenza vaccines: systematic review. *BMJ*, *338*, b354.

Kleinert, S., & Horton, R. (2014). How should medical science change?. *The Lancet*, *383*(9913), 197-198.

Macleod, M. R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J. P., ... & Glasziou, P. (2014). Biomedical research: increasing value, reducing waste. *The Lancet*, *383*(9912), 101-104.

Marco, C. A., & Larkin, G. L. (2000). Research ethics: ethical issues of data reporting and the quest for authenticity. *Academic Emergency Medicine*, *7*(6), 691-694.

Martin, B. R. (2013). Whither research integrity? Plagiarism , self-plagiarism and coercive citation in an age of research assessment. *Elsevier*, *42*(5), 1005–1014. https://doi.org/10.1016/j.respol.2013.03.011

Martinson, B. C., Thrush, C. R., & Crain, A. L. (2013). Development and validation of the survey of organizational research climate (SORC). *Science and engineering ethics*, *19*(3), 813-834.

Marusic, A., Wager, E., Utrobicic, A., Rothstein, H. R., & Sambunjak, D. (2016). Interventions to prevent misconduct and promote integrity in research and publication. *Cochrane Databasae of Systematic Reviews*, (4). https://doi.org/10.1002/14651858.MR000038.pub2

Mazar, N., & Ariely, D. (2015). Dishonesty in scientific research. *The Journal of clinical investigation*, *125*(11), 3993-3996.

Mondragón Barrios, L., Jiménez Tapia, J. A., Meza Mercado, D. M., & Sosa Mora, L. (2017). Regulation and self-regulation of ethical practices in scientific publication. *SALUD MENTAL*, *40*(5). JOUR. https://doi.org/10.17711/SM.0185-3325.2017.029

Olesen, A. P., Amin, L., & Mahadi, Z. (2017). Malaysian researchers talk about the influence of culture on research misconduct in higher learning institutions. *Accountability in Research*, *24*(8), 469–482. https://doi.org/10.1080/08989621.2017.1399358

Oliveira, G. S. De, Jung, M. J., & Mccarthy, R. J. (2015). Discrepancies Between Randomized Controlled Trial Registry Entries and Content of Corresponding Manuscripts Reported in Anesthesiology Journals. *Ovid Insights*, *121*(4), 1030–1033. https://doi.org/10.1213/ANE.0000000000000824

Sharma, O. P. (2015). Ethics in Science. *Indian Journal of Microbiology*, *55*(3), 341–344. https://doi.org/10.1007/s12088-015-0532-x

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384.

Steen, R. G., Casadevall, A., & Fang, F. C. (2013). Why Has the Number of Scientific Retractions Increased ? *PLoS ONE*, *8*(7), 1–9. https://doi.org/10.1371/journal.pone.0068397

Suhrbier, A., & Poland, G. A. (2013). Are Impact Factors corrupting truth and utility in biomedical research? *Elsevier*, *Vaccine 31*, 6041–6042.

I Tijdink, J. K., Schipper, K., Bouter, L. M., Pont, P. M., De Jonge, J., & Smulders, Y. M. (2016). How do scientists perceive the current publication culture? A qualitative focus group interview study among Dutch biomedical researchers. *BMJ open*, *6*(2), e008681.

II Tijdink, J. K., Bouter, L. M., Veldkamp, C. L. S., Ven, P. M. Van De, Wicherts, M., & Smulders, Y. M. (2016). Personality Traits Are Associated with Research Misbehavior in Dutch Scientists: A Cross-Sectional Study. *PLoS One*, 1–12. https://doi.org/10.1371/journal.pone.0163251

Tijdink, J. K. (2017). Publish & Perish; research on research and researchers. *Tijdschrift voor psychiatrie*, *59*(7), 406–413. English Abstract, Journal Article.

Yavchitz, A., Boutron, I., Bafeta, A., Marroun, I., Charles, P., Mantz, J., & Ravaud, P. (2012). Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. *PLoS medicine*, *9*(9), e1001308.

Wang, J., Ku, J. C., Alotaibi, N. M., & Rutka, J. T. (2017). Literature Review Retraction of Neurosurgical Publications: A Systematic Review. *World Neurosurgery*, *103*, 809–814.e1. https://doi.org/10.1016/j.wneu.2017.04.014

Zee, S. Van Der, Anderson, R., Poppe, R., & Jordan, J. (2016). When Lying Feels the Right Thing to Do. *Frontiers in Psychology*, *7*(June), 1–13.

SUPPLEMENTARY MATERIAL 2

Survey on publication practices in health services research

Attitude towards publication practices	**On the following page different statements regarding publication practices are presented. Please provide your opinion on the following statements. This section contains one page with statements.**	**Answer option likert scale** **1 = Strongly disagree** **2 = Disagree** **3 = Neither agree nor disagree** **4 = Agree** **5 = Strongly agree**
1	Please note that the following statements concern peer-reviewed Health Services Research specifically.	
2	A catchy title is necessary to get a manuscript published	
3	An abstract should be exciting to attract attention	
4	It is acceptable to describe conclusions beyond the research questions as long as you stay close to your results.	
5	It is acceptable if conclusions deviate slightly from the results as long as that is transparent.	
6	Recommendations for policy or practice should always be included in an international peer-reviewed publication.	
7	Generalizing findings outside the original context is acceptable to get a manuscript published.	
8	It is necessary to compare every single result to supporting and contradictory literature in the discussion section.	
9	In the limitations section, it is necessary to discuss potential impact of limitations on the findings.	
10	It is acceptable to make concluding statements on causality, even when a study design does not allow for inferences on causation.	
11	It is acceptable to use superlatives including ‘striking’, ‘impressive’, and ‘incredibly’, in the discussion section	
12	It is acceptable to describe conclusions more powerfully or stronger in public communication about the research (e.g. press releases, news items, LinkedIn messages) than in peer-reviewed publications.	
	Do you have any additional comments regarding the above statements?	[open answer]
Individual Researcher level	**The following section is about your individual researcher characteristics. Please indicate your agreement with the following statements. This section contains one page with statements.**	
1	In January 2016, I was employed as project leader or principal investigator.	
2	Please indicate your agreement with the following statements.	
3	I want to advance my career in science.	
4	To advance my career, it is important to produce many scientific publications in international peer-reviewed journals.	

5	To advance my career, it is important to have societal impact through e.g. television interviews, media exposure.	
6	It is important that my co-workers have a high opinion of my work.	
7	When my last manuscript was accepted by a peer-reviewed journal I was extremely happy.	
8	Before I start a research project, I have a clear idea of what the results will be.	
9	I hope to be surprised by my results.	
10	I am disappointed if my results are statistically not significant.	
11	I am sufficiently trained in the research methods I apply.	
12	I have received sufficient training in writing scientific publications.	
13	I find it difficult to write conclusions based on my research findings for peer-reviewed publications.	
14	I find it difficult to write recommendations for policy or practice in peer-reviewed publications.	
15	I have received sufficient training in research integrity.	
16	I have difficulty communicating with my co-authors about the contents of a manuscript.	
17	I get nervous when I receive feedback about my manuscript from my co-authors.	
18	Generally, when I receive a peer-review from a journal, I accept suggestions even when I do not entirely agree.	
19	Most of my peer-reviewed publications are excellently written.	
20	My peer-reviewed publications make an important contribution to my scientific field.	
	Do you have any additional comments regarding the above statements?	[open answer]
23 Junior	Generally, when writing a manuscript, I accept suggestions from my supervisors/PI/project leader even if I don't entirely agree.	
23 Senior	Generally, when supervising the writing of a manuscript, I make the final decisions on what text to include in the manuscript.	
Institutional level	**The following section is about the institute where you worked in January 2016. Please indicate your agreement with the following statements keeping the institute where you worked in January 2016 in mind. This section contains three pages with statements.**	
1	The institute where I worked provided training (opportunities) regarding the writing of discussion and conclusion sections specifically.	
2	The institute where I worked provided sufficient training on presenting in writing or verbally my findings in lay terms.	
3	The institute where I worked rewards high quality publications.	
4	The institute where I worked rewards high numbers of publications.	
5	The institute where I worked stimulates high quality publications as opposed to a high quantity of publications.	
6	The institute where I worked kept track of my societal impact.	

7	I was aware of the research code of conduct at the institute where I worked and understand its contents.	
8	Before I was hired at the institute where I worked, I was asked to demonstrate my writing ability.	
9	I received training on the formal quality assurance policy regarding the conduct of research at the institute where I worked.	
10	Do you have any additional comments regarding the above statements?	
11	I received feedback on my manuscript from my co-workers (not co-authors).	
12	I find it helpful to receive feedback on my manuscript from my co-workers.	
13	It was mandatory to discuss manuscripts during formal peer-review groups at the institute where I worked.	
14	In formal peer-review groups, suggestions were made for the revision in the discussion and conclusion sections of manuscripts.	
15	At the institute where I worked, I always discussed my manuscripts in a voluntary peer-review group.	
16	It was compulsory to address the review comments received in the peer-review group.	
17	There is a strong culture of competitiveness within the department/institute where I worked.	
18	I preferred not to share my findings with my co-workers before they were published.	
19	At the institute where I worked, ideas for new research studies (proposals) are discussed amongst colleagues.	
20	At the institute where I worked, I regularly attended a journal club together with my co-workers.	
21	At the institute where I worked, I regularly consulted a native English speaker on the writing of my manuscripts.	
22	At the institute where I worked, I could easily approach a statistician for help with the interpretation of my findings.	
23	When writing a public communication (e.g. press release, executive summary), I was always assisted by a communication expert.	
24 junior	My supervisor/PI/project leader at the institute where I worked did not have enough time to look at my manuscripts.	
25 Junior	My supervisor/PI/project leader at the institute where I worked has strong social skills.	
26 Junior	At the institute where I worked, I followed a course in academic writing before January 2016.	
24 Senior	I do not have sufficient time review manuscripts my juniors/PhD students write.	
25 Senior	I have strong social skills.	
26	Who bears final responsibility over the final version of a manuscript and its contents?	1 = First author / 2 = Last author / 3 = All authors / 4 = The institute - department

	Do you have any additional comments regarding the above statements?	[open answer]
27	The institute where I worked stimulates data sharing and open access policies.	
28	The institute where I worked has strict rules on the storage of data and filing of research material.	
29	It is common for the institute where I worked to write a press release when research results are published.	
30	t the institute where I worked, I frequently wrote a public communication (e.g. report, factsheet, policy brief) on my research findings aimed at policy or practice, separately from peer-reviewed scientific publication.	
31	It was easy to discuss any problems regarding my research with my co-workers.	
32	The institute where I worked actively stimulated informal interaction between co-workers.	
33	Generally, I had enough time to sit down and write a good manuscript.	
34	I experience high work pressure.	
35	I needed to report innovative and novel conclusions in my scientific work to obtain new funding.	
36	At the institute where I worked, funders made requests regarding the phrasing of conclusions and messages of my manuscripts.	
37	At the institute where I worked, I adapted my discussion or conclusion upon request of the funding agency at least once.	
38	When writing my manuscript, I felt I need to take the position of my funder into account.	
	Do you have any additional comments regarding the above statements?	[open answer]
Research Environment Level	**The following section is about research environment characteristics. Please indicate your agreement with the following statements. This section contains the last two pages with statements.**	
1	When I frame my conclusions more excitingly, I receive more citations.	
2	To make a career in science, I need to demonstrate societal impact.	
3	I sometimes approach journalists or other media to achieve public exposure of my results.	
4	Usually, I summarize my conclusions in social media such as twitter, LinkedIn or Facebook.	
5	My publications should generate media attention.	
6	Exciting conclusions will generate media attention.	
7	Journalists have exaggerated conclusions of my study in a public communication (i.e. press releases, news items) at least once.	
8	There is a lot of competition for better research positions in my field.	
9	My publication track record is essential to compete for better research positions.	

10	I feel pressure to publish my research in high impact journals.	
11	Generally, I try to publish in journals with the highest impact factors.	
12	I experience a high publication pressure.	
13	Without publication pressure, my scientific output would be of higher quality.	
14	I suspect that for some co-workers, publication pressure leads to inappropriate (e.g. exaggerated) conclusions or messages.	
	Do you have any comments regarding the above statements?	[open answer]
15	Journals will not accept my manuscript unless I frame 'exciting' conclusions or messages.	
16	It has happened that in the editing process at a journal, parts of the discussion or conclusions in my manuscript were significantly changed by the editor.	
17	My manuscripts would not have been accepted if I would not have emphasized 'positive' findings.	
18	When writing my manuscripts I always use a reporting checklist (e.g. STROBE, CONSORT), also when the journal does not require one.	
19	The word limit of journals hampers me from writing a good discussion and conclusions section in my manuscripts.	
20	I have experienced that one of my co-authors had a conflict of interest with regard to the research findings.	
21	I have experienced that a co-author pressured me to write conclusions that suited their own practice.	
22	I have experienced that there was a disturbing conflict between co-authors about the content of a manuscript.	
23	My publications are of better quality if fewer co-authors are involved.	
24	My publications are of better quality if co-authors from other institutes are involved.	
25	In the design of my study, I frequently involve stakeholders such as patients or professionals (not funders).	
26	I always discuss my preliminary findings and conclusions with stakeholders such as patients or professionals (not funders).	
	Do you have any additional comments regarding the above statements?	[open answer]
General information	What is your current (research) position?	[list of positions]
	Please specify,	
	How many years had you been working as a researcher in January 2018?	
	What was your research position in January 2016	[list of positions]
	Please specify,	

	In which academic background were you trained?	1 = Social Sciences 2 = Epidemiology 3 = (Health) economics 4 = Other
	Please specify,	
	What is your area of expertise?	[open question]
	How many articles have you currently co-authored in peer reviewed journals?	1 = 1 2= 2-5 3= 6-10 [.....]
	What is your age?	
End of survey		

SUPPLEMENTARY MATERIAL 3

Additional information on the exploratory factor analyses

This document describes the methods and results of the exploratory factor analyses conducted to identify factor domains from the survey 'Publication Practices in Health Services Research'.

Methods

The survey 'Publication Practices in Health Services Research' contains three domains: the individual domain existing of 18 items, the institutional domain existing of 34 items, and the scientific environmental domain existing of 26 items. We applied an exploratory factor analysis using principal component analysis as extracting method and direct oblimin as rotation method. We used these methods because factors were unlikely to be fully independent from each other. An average score for each factor was used to allow for an equal 'weighting' among the factors in the statistical analysis. Items with a load factor 0.3 or above were assigned into a specific factor. Items with multiple load factors were assigned to the most fitting factor based on the largest value of load factors on this specific item. A specific label was given to each particular factor. Based on the researchers' judgement, items which did not fit the assigned factor were removed. A reliability analyses was performed.

From the exploratory factor analysis, we identified 6 factors for the individual domain, 7 factors for institutional domain, and 7 factors for the scientific environmental domain. We conducted a reliability analysis for each factor by using Cronbach's alpha. Items with a Cronbach's alpha over 0.3 were considered for the bivariate analyses.

Results

Table S2.1. describes the factors resulting from the exploratory factor analyses. Table S2.2 describes the reliability of the resulting factors. Two items that strongly decreased the reliability of their assigned factors were removed. Four items that were categorised together by the factor analyses did not relate well conceptually i.e. items 'Individual_3, Individual_18, Environment_two_8, and Environment_two_6. These items were therefore considered as single-item factors.

Table S2.1 | Domains identified from the exploratory factor analysis.

FACTOR	ITEM NUMBER	FACTOR LOADING	ITEM DESCRIPTION
Domain: Individual			
Ambition in science	Individual_1	0.679	I want to advance my career in science.
	Individual_2	0.574	To advance my career, it is important to produce many scientific publications in international peer-reviewed journals.
	Individual_4	0.640	It is important that my co-workers have a high opinion of my work.
	Individual_8	0.476	I am disappointed if my results are statistically not significant.
Pressure to create societal impact	Individual_3	-	To advance my career, it is important to have societal impact through e.g. television interviews, media exposure.
Contribution to science	Individual_18	-	My peer-reviewed publications make an important contribution to my scientific field.
Self-efficacy	Individual_5	0.658	When my last manuscript was accepted by a peer-reviewed journal I was extremely happy.
	Individual_14	0.640	I have difficulty communicating with my co-authors about the contents of a manuscript.
	Individual_15	0.758	I get nervous when I receive feedback about my manuscript from my co-authors.
Perception of received training	Individual_7	0.546	I hope to be surprised by my results.
	Individual_10	0.617	I have received sufficient training in writing scientific publications.
	Individual_13	0.775	I have received sufficient training in research integrity.
Confidence in writing	Individual_9	0.345	I am sufficiently trained in the research methods I apply.

Table S2.1 | Continued

FACTOR	ITEM NUMBER	FACTOR LOADING	ITEM DESCRIPTION
	Individual_11	0.623	I find it difficult to write conclusions based on my research findings for peer-reviewed publications.
	Individual_12	0.628	I find it difficult to write recommendations for policy or practice in peer-reviewed publications.
Domain: Institutional			
Specific education	InstitutionalOne_1	0.699	The institute where I worked provided training (opportunities) regarding the writing of discussion and conclusion sections specifically.
	InstitutionalOne_2	0.767	The institute where I worked provided sufficient training on presenting in writing or verbally my findings in lay terms.
Competitiveness	InstitutionalOne_3	0.758	The institute where I worked rewards high quality publications.
	InstitutionalOne_4	0.791	The institute where I worked rewards high numbers of publications.
	InstitutionalTwo_6	0.663	There is a strong culture of competitiveness within the department/institute where I worked.
Data storage	InstitutionalOne_9	0.492	I received training on the formal quality assurance policy regarding the conduct of research at the institute where I worked.
	InstitutionalThree_2	0.750	The institute where I worked has strict rules on the storage of data and filing of research material.
Feedback culture at institute	InstitutionalTwo_1	0.882	I received feedback on my manuscript from my co-workers (not co-authors).
	InstitutionalTwo_2	0.842	I find it helpful to receive feedback on my manuscript from my co-workers.
	InstitutionalTwo_3	0.807	It was mandatory to discuss manuscripts during formal peer-review groups at the institute where I worked.
	InstitutionalTwo_4	0.829	In formal peer-review groups, suggestions were made for the revision in the discussion and conclusion sections of manuscripts.

Table S2.1 | Continued

FACTOR	ITEM NUMBER	FACTOR LOADING	ITEM DESCRIPTION
	InstitutionalTwo_10	0.684	At the institute where I worked, I always discussed my manuscripts in a voluntary peer-review group.
	InstitutionalTwo_5	0.408	It was compulsory to address the review comments received in the peer-review group.
Social support	InstitutionalTwo_8	0.415	At the institute where I worked, ideas for new research studies (proposals) are discussed amongst colleagues.
	InstitutionalThree_5	0.575	It was easy to discuss any problems regarding my research with my co-workers.
	InstitutionalThree_6	0.727	The institute where I worked actively stimulated informal interaction between co-workers.
	InstitutionalThree_7	0.778	Generally, I had enough time to sit down and write a good manuscript.
	InstitutionalThree_8	0.387	I experience high work pressure.
Media policy	InstitutionalThree_3	0.690	It is common for the institute where I worked to write a press release when research results are published.
	InstitutionalThree_4	0.698	At the institute where I worked, I frequently wrote a public communication (e.g. report, factsheet, policy brief) on my research findings aimed at policy or practice, separately from peer-reviewed scientific publication.
	InstitutionalThree_9	0.308	I needed to report innovative and novel conclusions in my scientific work to obtain new funding.
Influence of funders	InstitutionalThree_10	0.751	At the institute where I worked, funders made requests regarding the phrasing of conclusions and messages of my manuscripts.
	InstitutionalThree_11	0.813	At the institute where I worked, I adapted my discussion or conclusion upon request of the funding agency at least once.
	InstitutionalThree_12	0.809	When writing my manuscript, I felt I need to take the position of my funder into account.

Table S2.1 | Continued

FACTOR	ITEM NUMBER	FACTOR LOADING	ITEM DESCRIPTION
Domain: Environment			
Creating exciting conclusion	ResearchEnvironment_1	0.797	When I frame my conclusions more excitingly, I receive more citations.
	EnvironmentTwo_1	0.736	Journals will not accept my manuscript unless I frame 'exciting' conclusions or messages.
Media contact	ResearchEnvironment_3	0.601	I sometimes approach journalists or other media to achieve public exposure of my results.
	ResearchEnvironment_4	0.799	Usually, I summarize my conclusions in social media such as twitter, LinkedIn or Facebook.
	ResearchEnvironment_5	0.496	My publications should generate media attention.
	ResearchEnvironment_7	0.632	Journalists have exaggerated conclusions of my study in a public communication (i.e. press releases, news items) at least once.
Pressure from scientific culture	ResearchEnvironment_8	0.781	There is a lot of competition for better research positions in my field.
	ResearchEnvironment_9	0.627	My publication track record is essential to compete for better research positions.
	ResearchEnvironment_10	0.853	I feel pressure to publish my research in high impact journals.
	ResearchEnvironment_11	0.717	Generally, I try to publish in journals with the highest impact factors.
	ResearchEnvironment_12	0.783	I experience a high publication pressure.
Suspicion of co-workers	ResearchEnvironment_13	0.599	Without publication pressure, my scientific output would be of higher quality.
	ResearchEnvironment_14	0.516	I suspect that for some co-workers, publication pressure leads to inappropriate (e.g. exaggerated) conclusions or messages.
	EnvironmentTwo_7	0.807	I have experienced that a co-author pressured me to write conclusions that suited their own practice.
	EnvironmentTwo_9	0.752	My publications are of better quality if fewer co-authors are involved.

Table S2.1 | Continued

FACTOR	ITEM NUMBER	FACTOR LOADING	ITEM DESCRIPTION
Journal practice	EnvironmentTwo_2	0.572	It has happened that in the editing process at a journal, parts of the discussion or conclusions in my manuscript were significantly changed by the editor.
	EnvironmentTwo_5	0.757	The word limit of journals hampers me from writing a good discussion and conclusions section in my manuscripts.
Stakeholder influence	EnvironmentTwo_11	0.784	In the design of my study, I frequently involve stakeholders such as patients or professionals (not funders).
	EnvironmentTwo_12	0.879	I always discuss my preliminary findings and conclusions with stakeholders such as patients or professionals (not funders).
Co-author conflict of interest	EnvironmentTwo_6	-	I have experienced that one of my co-authors had a conflict of interest with regard to the research findings.
Disturbing conflict between co-authors	EnvironmentTwo_8	-	I have experienced that there was a disturbing conflict between co-authors about the content of a manuscript.

Table S2.2 | Factors score and the results from reliability analysis

Factors	Domains	Mean	SD	Cronbach's alpha
Individual	Ambition in science	3.53	0.52	0.46
	Self-efficacy	2.36	0.59	0.57
	Perception of received training	3.66	0.57	0.51
	Confidence in writing	2.85	0.57	0.42
	Pressure to create social impact	3.41	0.78	n.a
	Perception of contribution to science	3.71	0.69	n.a
Institution	Specific training	3.19	0.95	0.73
	Competitiveness	3.40	0.90	0.73
	Data storage	3.47	0.86	0.37
	Feedback culture at institute	3.17	0.91	0.85
	Social support	3.68	0.55	0.52
	Media policy	2.68	0.79	0.54
	Influence of funders	1.74	0.73	0.79
Environment	Creating exciting conclusion	2.69	0.67	0.52
	Media contact	2.46	0.78	0.63
	Pressure from scientific culture	3.78	0.66	0.82
	Suspicions of co-workers	2.39	0.66	0.69
	Journal practice	2.28	0.70	0.34
	Stakeholder influence	3.13	1.01	0.71
	Co-author conflict of interest	1.87	0.84	n.a
	Conflict between co-author	2.08	1.01	n.a

n.a: not applicable, single item question

SUPPLEMENTARY MATERIAL 4

In this document the bivariate analyses and multivariate models for the stratification between first and last authors are provided.

Table S4.1 | Bivariate analysis between factors from individual, institutional, and scientific environment domain with number of QRPs using Poisson regression (first and last author)

Domain	Factors	First author					Last author				
		B	SE	p-value	Exp(B)	95% CI	B	SE	p-value	Exp(B)	95% CI
Individual	Ambition in science	0.141	0.116	0.223	1.15	0.92-1.45	-0.025	0.145	0.862	0.98	0.73-1.30
	Self-efficacy	-0.060	0.100	0.552	0.94	0.77-1.15	-0.381	0.139	0.006	0.68	0.52-0.90
	Perception of received training	-0.173	0.105	0.099	0.84	0.69-1.03	0.223	0.140	0.112	1.25	0.95-1.64
	Confidence in writing	-0.008	0.110	0.992	0.99	0.82-1.23	-0.268	0.151	0.077	0.76	0.57-1.03
	Pressure to create societal impact	0.315	0.089	0.001	1.37	1.15-1.63	0.256	0.093	0.006	1.29	1.08-1.55
	Perception of contribution to science.	0.112	0.095	0.240	1.12	0.93-1.35	0.071	0.108	0.512	1.07	0.87-1.32
Institution	Specific training in reporting messages and conclusions	-0.199	0.059	0.001	0.82	0.73-0.92	-0.108	0.082	0.189	0.90	0.76-1.05
	Competitiveness	0.054	0.074	0.463	1.06	0.91-1.22	0.119	0.080	0.139	1.13	0.96-1.32
	Data storage	-0.097	0.068	0.154	0.91	0.79-1.03	0.032	0.089	0.722	1.03	0.87-1.23
	Feedback culture at institute	0.117	0.069	0.091	1.13	0.98-1.28	-0.090	0.076	0.240	0.91	0.78-1.06
	Social support	-0.046	0.103	0.653	0.95	0.78-1.17	-0.020	0.142	0.887	0.98	0.74-1.29

Table S4.1 | Continued

Domain	Factors	First author					Last author				
		B	SE	p-value	Exp(B)	95% CI	B	SE	p-value	Exp(B)	95% CI
	Media policy	0.090	0.095	0.346	1.09	0.91–1.32	-0.038	0.075	0.612	0.96	0.83–1.16
	Influence of funders	-0.159	0.085	0.061	0.85	0.72–1.01	0.084	0.110	0.449	1.09	0.88–1.35
Environment	Creating exciting conclusion	0.280	0.120	0.020	1.32	1.04–1.68	-0.111	0.087	0.203	0.89	0.75–1.06
	Media contact	0.005	0.082	0.954	1.01	0.86–1.18	-0.090	0.098	0.360	0.91	0.76–1.11
	Pressure from scientific culture	0.028	0.105	0.789	1.03	0.84–1.26	0.047	0.099	0.634	1.05	0.86–1.27
	Suspicions of co-workers	-0.058	0.087	0.507	0.95	0.79–1.12	-0.182	0.124	0.143	0.83	0.65–1.06
	Journal practice	-0.107	0.087	0.221	0.90	0.76–1.07	-0.063	0.105	0.550	0.94	0.76–1.15
	Stakeholder influence	0.107	0.061	0.080	1.11	0.98–1.25	0.196	0.073	0.007	1.22	1.05–1.40
	Co-author conflict of interest	-0.024	0.0794	0.758	0.98	0.84–1.14	-0.207	0.085	0.016	0.81	0.68–0.96
	Conflict between co-authors	-0.015	0.0611	0.801	0.98	0.87–1.11	-0.152	0.066	0.023	0.86	0.75–0.98

Table S4.2 | Multivariate analysis between factors from individual, institutional, and scientific environment domains with number of QRPs using Poisson regression (first author).

	Model 1[a]			Model 2[b]			Model 3[a]			Model 4[b]		
	B	Exp(B)	95% CI	B	Exp(B)	95% CI	B	Exp(B)	95% CI	B	Exp(B)	95% CI
Intercept	0.016	1.02	0.15-7.09	1.180	3.25	1.10-9.62	0.248	1.28	0.18-0.93	1.261	3.53	1.06-11.67
Journal impact factor	-	-	-	-	-	-	-0.114	0.89	0.80-1.01	-0.076	0.93	0.84-1.02
Working duration	-	-	-	-	-	-	0.004	1.01	0.98-1.03	0.012	1.01	0.99-1.03
Individual												
Ambition in science	0.271	1.31	0.99-1.72	-	-	-	0.306	1.36	0.99-1.85	-	-	-
Perception of received training	-0.142	0.87	0.68-1.11	-	-	-	-0.139	0.87	0.68-1.14	-	-	-
Pressure to create societal impact	0.257	1.29	1.07-1.56	0.279	1.32	1.11-1.58	0.216	1.24	1.02-1.51	0.247	1.28	1.07-1.54
Perception of contribution to science.	0.139	1.20	0.93-1.53	-	-	-	0.192	1.21	0.94-1.55	-	-	-
Institution												
Specific training	-0.185	0.83	0.71-0.96	-0.165	0.85	0.75-0.96	-0.170	0.84	0.72-0.98	-0.174	0.84	0.74-0.96
Data storage	-0.018	0.98	0.82-1.17	-	-	-	-0.112	0.89	0.73-1.10	-	-	-
Feedback culture	0.158	1.17	1.01-1.36	-	-	-	0.218	1.24	1.05-1.47	-	-	-
Influence of funders	-0.188	0.83	0.69-0.99	-	-	-	-0.173	0.84	0.70-1.02	-	-	-
Environment												
Creating exciting conclusion	0.075	1.08	0.81-1.44	0.054	1.06	0.81-1.37	0.060	1.06	0.79-1.44	0.087	1.09	0.84-1.42
Journal practice	-0.090	0.91	0.76-1.10	-	-	-	-0.075	0.93	0.77-1.12	-	-	-
Stakeholder influence	0.060	1.06	0.92-1.23	-	-	-	0.052	1.05	0.90-1.24	-	-	-

[a] *Included domains with p-value <0.30 in bivariate analysis;* [b] *Included domains with p-value <0.05 in bivariate analysis;*

Table S4.3 | Multivariate analysis between factors from individual, institutional, and scientific environment domains with number of QRPs using Poisson regression (last author).

	Model 1[a]			Model 2[b]			Model 3[a]			Model 4[b]		
	B	Exp(B)	95% CI	B	Exp(B)	95% CI	B	Exp(B)	95% CI	B	Exp(B)	95% CI
Intercept	2.201	9.03	1.54-52.89	2.034	7.65	2.48-23.57	2.212	9.13	1.47-56.60	2.132	8.43	2.56-27.74
Journal impact factor	-	-	-	-	-	-	-0.066	0.94	0.86-1.02	-0.066	0.94	0.87-.1.01
Working duration	-	-	-	-	-	-	0.005	1.01	0.98-1.02	0.004	1.01	0.98-1.02
Individual												
Self-efficacy	-0.190	0.83	0.59-1.16	-0.304	0.74	0.56-0.98	-0.198	0.82	0.59-1.15	-0.308	0.74	0.55-0.98
Perception of received training	0.266	1.30	0.96-1.77	-	-	-	0.253	1.29	0.95-1.75	-	-	-
Confidence in writing	-0.359	0.78	0.48-1.01	-	-	-	-0.360	0.70	0.48-1.01	-	-	-
Pressure to create societal impact	0.215	1.24	0.93-1.65	0.249	1.28	1.01-1.62	0.175	1.19	0.90-1.58	0.173	1.19	0.93-1.52
Institution												
Specific training	-0.114	0.89	0.74-1.08	-	-	-	-0.058	0.94	0.77-1.16	-	-	-
Competitiveness	0.111	1.12	0.89-1.40	-	-	-	0.099	1.10	0.88-1.39	-	-	-
Feedback culture	-0.068	0.93	0.77-1.13	-	-	-	-0.094	0.91	0.75-1.10	-	-	-
Environment												
Creating exciting conclusion	0.003	1.01	0.78-1.28	-	-	-	-0.052	0.95	0.73-1.24	-	-	-
Suspicion of co-workers	-0.031	0.97	0.72-1.31	-	-	-	0.047	1.05	0.76-1.44	-	-	-
Stakeholder influence	-0.043	0.96	0.75-1.22	-0.014	0.98	0.80-1.21	-0.012	0.98	0.77-1.27	0.037	1.04	0.84-1.29
Co-author conflict of interest	-0.112	0.89	0.74-1.08	0.159	0.85	0.71-1.03	-0.083	0.92	0.75-1.12	-0.122	0.88	0.73-1.08
Conflict between co-authors	-0.108	0.90	0.75-1.07	-0.036	0.96	0.82-1.13	-0.118	0.89	0.74-1.06	-0.052	0.95	0.81-1.12

[a] *Included domains with p-value <0.30 in bivariate analysis;* [b]*Included domains with p-value <0.05 in bivariate analysis;*

8

General Discussion

This thesis addresses two topics. First, it explores the interpretation of publicly reported statistics and proposes a method to improve the interpretation of publicly reported statistics on health and healthcare. Second, the thesis explores the responsible reporting of health services research (HSR) in scientific and societal publications. This final chapter discusses the methodological considerations of the research and reflects on the findings of the included studies, with practical recommendations for research funders, HSR funders, leaders of HSR institutions, scientific journal editors, individual researchers, science communicators/journalists, and decision-makers: the users of the research findings (i.e. policy makers, healthcare professionals, the general public and patients).

This thesis addressed multiple research questions by applying a variety of data sources and methodologies, as summarized in table 1 in the introduction.

MAIN FINDINGS

To aid HSR funders, research institutions, researchers, scientific journals, science communicators/journalists, policy makers, healthcare professionals, and the general public in the responsible interpretation and reporting of healthcare statistics and HSR, the aims of this thesis were to 1) propose a method to improve the interpretation of publicly reported statistics on health and healthcare and 2) provide insight into the scientific and public reporting of Health Services Research. To address the first aim, this thesis describes the use of statistics in the policy debates, the key characteristics relevant to the interpretation of publicly reported statistics, and the development of a tool to increase the likelihood of accurate interpretation of statistics on health and healthcare. To address the second aim, this thesis described the occurrence and nature of questionable research practices (QRPs) in the reporting of messages and conclusions in scientific HSR publications, reporting inadequacies and inconsistencies in societal HSR publications, and what factors may be associated with QRPs in scientific HSR publications.

Table 1 | Overview of chapters and respective study characteristics

Chapter	Research questions	Data source	Methodology
Part I			
2	How are different types of statistics used in various ongoing policy debates on health and healthcare in the Netherlands?	Policy documents on 8 topics derived from the Dutch parliament and government websites	Qualitative, conventional content analysis
3	(1) What are the key characteristics relevant to the interpretation of statistics on health and healthcare? (2) What method can improve the proper use of these statistics by policymakers, managers, scientists, patients, and the general public?	Expert consultation, focus groups, semi-structured interviews	Qualitative, thematic content analyses
4	(1) What is the usability of the FIAT-Health 1.0 amongst its intended user groups? (2) How can the FIAT-Health 1.0 be improved?	Test and evaluation forms of FIAT-Health 1.0, expert consultation, test session FIAT-Health 2.0	Qualitative, content analyses
Part II			
5	(1) What is the occurrence and nature of Questionable Research Practices (QRPs) in the reporting of messages and conclusions in international scientific Health Service Research (HSR) publications authored by researchers from HSR institutions in the Netherlands? (2) What is the relationship between study type, methodology and design and the occurrence of QRPs?	International scientific HSR publications authored by researchers from HSR institutions in the Netherlands	Mixed methods, assessment of scientific publications, quantitative analyses of relationship study type. methodology, and methodology
6	(1) Are societal publications on HSR consistent with the messages reported in the underlying research papers? (2) Are QRPs in scientific HSR publications replicated in societal publications? (3) Do fewer inconsistencies occur in societal publications if the first scientific author is involved in writing them?	International scientific HSR publications, and societal HSR publications that derived from them	Mixed methods, Qualitative content analysis of replicated QRPs and inconsistencies in societal publications Quantitative analyses of contribution of scientific author
7	How do individual, institutional, and scientific environment factors associate with QRPs messages and conclusions in scientific HSR publications?	International scientific HSR publications, literature review, semi-structured interviews, focus groups, and survey data	Mixed-methods, Explorative literature review, qualitative thematic analyses Quantitative analyses of survey data with assessed QRPs

Part I. Responsible Interpretation of Statistics on Health and Healthcare

Part I of this thesis addressed the interpretation of statistics on health and healthcare.

Chapter 2

Chapter 2 evaluates the use of statistics in the policy debate via an analyses of Dutch policy documents written between 2014 and 2016 on eight different topics on health and healthcare. We found that sources of the statistics used in policy debates were primarily government funded. We identified two distinct functions of statistics in these documents (i.e., rhetorical and managerial). When the specific statistic was used for agenda setting or to convince the public of the importance of a topic, the statistic was used rhetorically. Statistics that were used for planning, monitoring or evaluation of policy were used managerially. When evaluating a specific policy, applied statistics were mostly the result of routine or standardized data collection. Unlike rhetorical policy debates, managerial policy debates mirrored terms used by the scientific community. For instance, terms such as case-mix and statistical significance were used by policy makers to support their arguments.

Chapter 3

Chapter 3 describes the development of a tool to facilitate better interpretation of statistics on health and healthcare. We conducted focus groups and semi-structured interviews with experts on morbidity, healthcare expenditure, healthcare outcomes, and lifestyle statistics. Identified characteristics relevant to the interpretation were the statistics' origin, credibility, expression, subject matter, population and geographical focus, time period, and underlying data collection methods. Characteristics were translated into questions that were used to construct the Figure Interpretation Assessment Tool – Health (FIAT-Health) 1.0.

Chapter 4

Chapter 4 describes the testing and improvement of the FIAT-Health 1.0. Potential users assessed one publicly reported statistic on morbidity, healthcare expenditure, healthcare outcomes, or lifestyle and shared their experience in using the FIAT-Health. For each statistic, an expert on the specific topic provided a comparative assessment. Participating researchers, communication officers, and policy makers considered the questions of the FIAT-Health 1.0 useful. Expert assessments were comparable to the questions of the FIAT-Health. However, potential users reported that the format and language needed improvement. The tool was refined according to the results of the test and evaluation and transformed from a quantitative scoring instrument into an online qualitative appraisal tool: the FIAT-Health 2.0.

Part II. Responsible Reporting of Health services research

Part II of this thesis addressed the responsible reporting of Health Services Research in scientific and societal publications.

Chapter 5

Chapter 5 describes the occurrence and nature of QRPs in the reporting of messages and conclusions in scientific HSR publications. Together with 13 participating HSR institutions in the Netherlands, we constructed a definition of QRPs in the reporting of messages and conclusions in HSR and developed an assessment instrument covering 35 possible QRPs. Using the assessment instrument, we assessed a random sample of 116 HSR articles authored by researchers from the 13 participating HSR institutions published in international peer-reviewed scientific journals in 2016. A median of six QRPs per publication was identified. QRPs occurred most frequently in the reporting of implications for policy and practice, recommendations for policy and practice, contradictory evidence, study limitations, and conclusions based on the results and in the context of the literature. We identified no differences in the total number of QRPs in publications based on different methodological approaches, type of research, or study design.

Chapter 6

Chapter 6 addresses the QRPs in the reporting of messages, along with inconsistencies in the conclusions of societal HSR publications. We conducted a directed qualitative content analysis of societal publications derived from the scientific HSR publications assessed in chapter 5. Our results indicated that most societal publications contained some inconsistencies or replicated QRPs from the corresponding scientific publications. These issues were most often characterized by inadequately reported conclusions, policy and practice recommendations, and titles. No substantial differences emerged overall between societal publications produced by authors of the scientific publication, the research institutions of funders, and those written by outsiders.

Chapter 7

The study into factors related to QRPs in scientific publications is described in chapter 7. The research question was addressed through a mixed-methods study design. First, we identified factors possibly contributing to the occurrence of QRPs in scientific HSR publications through a literature review, 14 semi-structured interviews with leaders of HSR groups or institutions in the Netherlands, and 13 focus groups amongst junior health services researchers. Pressure to create societal impact was associated with a higher number of QRPs in the reporting of messages and conclusions in HSR publications. Specific training in reporting messages and conclusions, as well as awareness of co-author conflicts of interests, were related to fewer QRPs in HSR publications.

METHODOLOGICAL CONSIDERATIONS

Various methodological considerations need to be discussed that affected the inferences made in this thesis. Each chapter describes a number of limitations that were considered in respective discussion sections. A subsequent section describes overall methodological considerations and generalizability of the findings from this thesis. The following subjects are discussed: stakeholder involvement, selection of study documents and participants, the conceptualization of 'questionable research practices in the reporting of messages and conclusions', and the generalizability of the findings.

This thesis addressed topics that have not been previously studied in the Health Services Research literature and was therefore primarily explorative. This approach leaves ample room for methodological improvements, however, it also allowed us to make great strides in creating awareness amongst participants in the field of HSR on issues pertaining to the interpretation of statistics on health and healthcare and the scientific and societal reporting of HSR.

Stakeholder involvement

The studies in the thesis were carried out with substantive stakeholder involvement in the design and interpretation of findings. The first part of this thesis, "Responsible interpretation of statistics on health and healthcare", involved an advisory group and potential users of the FIAT-Health (i.e., researchers, science journalists/communicators, knowledge integration specialists who professionally engage in the dissemination of research findings towards policy makers or the public, and policy makers). The first part of this thesis addresses the interpretation of statistics. Potential users were involved in the construction and testing of the FIAT-Health 2.0, and students were involved as a proxy for the public. Through their involvement, participants actively increased their awareness of characteristics relevant to the interpretation of statistics and developed skills useful for their work.

The second part of this thesis, "Responsible reporting of health services research", was conducted in collaboration with 13 HSR institutions and departments from both academic and non-academic institutions in the

Netherlands. These institutions all agreed to participate in an effort to assure the overall quality of HSR publications in the Netherlands.

Active stakeholder involvement supported the quality of the research and the sustainability of the instruments and recommendations provided in this thesis. HSR institution leaders were involved in the design methodology and interpretation of research. This allowed the research community to develop HSR reporting norms and explicit methodologies. By doing so, the research community can continue to refine and improve these norms and methodologies in the future (1). Additionally, stakeholder involvement resulted in additional gains outside the scope of this thesis by facilitating broader discussions on research policies in the HSR community (2), such as developments in privacy law and graduate education.

Selection of study documents and participants

For the purposes of analyses, data were obtained in the form of documents (e.g., policy documents, scientific and societal publications), interviews with focus groups, and surveys amongst researchers, science communicators/ journalists, knowledge integration specialists, policy makers, and students.

Chapter 2 and 6 were entirely or partially grounded in document analyses. To collect these documents, systematic search strategies were used. Only publicly accessible online documents were included in our analyses. Any documents that were private, not published online, or deleted were excluded from analyses. Policy debate described in these documents may have led to different insights as compared to those that remained in the public domain. Chapter 5 was grounded in the selection of scientific HSR publications derived from 13 HSR institutions in the Netherlands. Because many of these institutions publish on a broad range of research fields, a common definition of HSR was used to identify HSR publications from publication lists. Although the inclusion of a publication was discussed within the full research group, it is possible that some HSR publications that partially fall under this definition may have been overlooked.

Chapter 3 and 4 relied on voluntary participation. Participants may have been aware of difficulties with statistical interpretation and were prepared to invest their time in joining the study. They might thus have believed in the usefulness of a tool for the assessment of publicly reported statistics beforehand, and might not criticize it completely. However, participants were professionals with different perspectives on the use of statistics. Each participant was therefore able to provide critical feedback to the construction, content and format of the tool.

The study described in chapter 7 was based on data from a sample of HSR researchers who either participated in focus groups or completed a survey. Participants who participated in the focus groups were encouraged by their institution to participate. Participants who completed the survey were randomly sampled from the full list of scientific HSR publications from 2016. Participants were mostly still affiliated as researchers at their respective HSR institutions. This may have resulted in a bias due to self-selection: those who stayed might have experienced a more positive research culture than those who left. Consequently, they might have been less affected by negative factors in that research culture. Moreover, in part II of this thesis, some quantitative findings are provided. Because of the scope and explorative character of this thesis, sample sizes in were generally small and our power to detect significant effects may have been modest. Yet, because our aim was explorative, the current sample sizes served the purpose of this thesis and pointed us in the direction of future research and areas for improvement.

Conceptualization of 'questionable research practices in the reporting of messages and conclusions'

In this thesis, 'questionable research practices in the reporting of messages and conclusions in HSR' is defined as *"to report, either intentionally or unintentionally, conclusions or messages that may lead to incorrect inferences and do not accurately reflect the objectives, the methodology or the results of the study."*

This definition was attained through a comprehensive process of literature review and interviews with directors/leaders of 13 HSR institutions. It was also validated through input from five leading international health services researchers. The definition is broad and incorporates normative judgement in distinguishing incorrect inferences from accurate reflection. Through consensus methods, 35 possible QRPs were identified that fall under this broad definition. QRPs such as unjustified causality, inadequately justified limitations, and conclusions that are not adequately based on the findings are generally accepted as questionable. Our definition additionally includes some QRPs that are debated amongst the HSR community, such as the inclusion of recommendations for policy and practice and the mention of contradictory evidence. The term 'questionable' is operationalized as reporting practices that 'raise questions'. These practices are therefore not necessarily incorrect, but under close inspection may be found lacking in clarity, transparency, or justification. When interpreting the findings presented in this thesis, the normative nature of our definition should be considered.

In addition, the construction of the concept involved consideration of a wide range of research designs, including qualitative designs. Studies that previously defined QRPs have primarily considered only quantitative research. Unlike QRPs in qualitative studies, quantitative QRPs are widely identified (e.g., cherry-picking, p-hacking, data-diving, overstating significance levels or effect sizes). This quantitative conceptualization of QRPs formed the basis of our understanding of what a QRP can encompass. Moreover, although numerous publication checklists exist for reporting quantitative studies of different designs, relatively few checklists exist describing responsible reporting of qualitative studies. This thesis sought to address this gap in the literature by fully incorporating qualitative research designs.

Generalizability of the findings

The data obtained for all studies in this thesis were derived from Dutch documents or obtained primarily from Dutch participants. The findings from each individual chapter are therefore to some extent restricted to the Dutch context. When discussing the use of statistics in the Dutch policy debate on health and healthcare, or the types of QRPs and inconsistencies identified in societal HSR literature, conclusions should thus only be applied to the Dutch context.

Nevertheless, most of the findings in this thesis are generalizable across the Dutch boarders.

Rationality for decision making differs for each country. The level to which evidence is considered valuable outside the Netherlands with a dissimilar tradition of evidence-based policy and practice will differ. The governance of healthcare institutions may affect the use of evidence and its ability to stimulate change. In countries where decision-making occurs through an integrated process with the collaboration of multiple institutions, like the Netherlands, evidence is more likely to stimulate a change of policy direction as opposed to countries with less integrated decision-making processes (3). The usability of a FIAT-Health would vary worldwide. Nevertheless, the construction of statistics on health and healthcare is subject to the same methodological limitations and interpretation challenges across the world. Additionally, the FIAT-Health was designed for the field of health and healthcare. However, it is likely that the characteristics addressed by the FIAT-Health 2.0 are transferable to statistics on different subjects, such as employment and social affairs. Consequently, the identified characteristics relevant to the interpretation of statistics and the international usability of the FIAT-Health 2.0 can likely be generalized to other countries and other subjects.

Further, the field of HSR is strongly connected internationally as the process of scientific publication is guided by international journals, including international peer-review, and the research is often performed by (inter) national consortia. Even though our findings were derived from the research published by 13 HSR institutions in the Netherlands, HSR publications from other Dutch HSR institutions, as well as institutions outside of the Netherlands, were represented in our sample. The occurrence of QRPs in messages and conclusions in scientific literature can therefore be generalized beyond the included 13 institutions and across the field of HSR.

Finally, the Netherlands has a relatively active HSR community. Because the community is well connected, it is possible to find consensus on measures to improve responsible reporting (1). Countries with less integrated HSR communities may be less inclined to act collectively to strengthen the quality of scientific reporting. Nonetheless, HSR institutions internationally can learn from the developments and factors identified in the Dutch context. Findings presented in part II of this thesis are also relevant knowledge for the HSR community internationally.

REFLECTION ON THE FINDINGS OF THIS THESIS

The reflection on the findings of this thesis are described in two parts. Part one will cover reflections on results pertaining to the Responsible Interpretation of Statistics on Health and Healthcare. Part two will include reflections on results pertaining to the Responsible Reporting of Health Services Research.

Responsible Interpretation of Statistics on Health and Healthcare

The findings of part I provide insight in the use of statistics in policy debate and the appraisal of publicly reported statistics on health and healthcare.

Use of statistics on health and healthcare in policy debate

The extent to which decision makers assess statistics, critically depends on their intentions. Our study presented in chapter 1 revealed a difference in the critical evaluation of statistics between rhetorical and managerial policy debates. Findings indicated policy makers and their political opponents did not allow for nuanced interpretation when using a statistic in rhetorical debate, while they did for managerial debates. Because the intention of rhetorical debate is to draw attention and set priorities for policy, a deep understanding of the statistic may not seem relevant or necessary to its user. After all, to draw attention, any report or speech requires a sweeping opening statement. A shining statistic serves this purpose well.

However, statistics in rhetorical debate should not be taken lightly. Statistics in the policy debate primarily function to provide 'insights into the nature of social problems' (4). Over time they provide the background for ideas and concepts used in policy decisions. Statistics used for rhetorical purposes settle in our common understanding and may change our values and priorities (4). Moreover, when statistics are used to determine the relative importance of social problems, the rationale of what is important is reduced to measurable aspects of healthcare while ignoring other important values that may be relevant to political debate (5). The appropriate use of and critical reflection on these statistics is therefore just as important as statistics that directly influence healthcare policy. For instance, in the Netherlands and internationally, the concept of Value Based Healthcare (VBHC) is increasingly used to financially reward care providers for their quality of care, and has additionally started to determine what value in healthcare means. In VBHC, value is defined as "the patient health outcomes achieved per dollar spent" (6). When healthcare value is debated by healthcare professionals and policy makers, they focus on the measured aspects of care that fall under the definition of VBHC. Aspects of valuable care not (yet) included in that definition, such as patients' unique personal concerns and expectations, the intrinsic value of caring, trust in professional accountability, and solidarity within the healthcare system, are subsequently omitted from professional and policy debate (7).

When statistics are used (or intended) for a managerial function, the debate within policy mirrors to a large extent the debate on these topics in the scientific community. When evaluating specific policy, statistics were primarily the result of routine or standardized data collection. The construction of these routine or standardized statistics was, however, not discussed in policy debate. Across the field of health and healthcare, we are increasingly reliant on standardized measurement and data use. Whether it be a national policy making setting, or a single doctor's consultation, there seems to be a movement towards a rational process of decision making informed by statistical information. Standardized and widely accepted measurement shapes a common understanding of a problem, creates insight into possible improvements, and allows a policy debate to focus on the actual problem at hand rather than the construction of the statistic itself. For instance, standardization of data collection on breast cancer has made it possible to track progress of health outcomes in breast cancer over the years. As is indicated by our findings, common trust and understanding of this statistic allows policy makers to debate the progress of care and evaluate if policy measures are necessary.

To facilitate the use of statistics to monitor, manage and evaluate healthcare, it appears worthwhile to invest in routine or standardized measurements. Nevertheless, a blind trust in statistics might also result in misuse. For instance, statistics on quality of care form the basis for the measurement and improvement of medical practices. Statistics for quality improvement take the form of quality indicators. Quality indicators are "explicitly defined and measurable items which act as building blocks in the assessment of care" (8). They are developed through rigorous methods and appear scientifically sound. Yet, quality indicators are regularly implemented in practice without critical appraisal of their usefulness (9). That is, quality indicators may not be adequately valid or reliable for improvement of care provider performance (10). They may be used to provide information on aspects of care outside the scope of their initial purpose (9), and may lack relevance in the perspective of other values, as the indicator only covers a limited scope of care provided (9). Consequently, if not approached critically, quality indicators might misinform the improvement of quality of care, as opposed to facilitating it.

Before any statistic can be used responsibly in policy or practice, decision makers should examine and discuss them critically. If statistics are questioned and negotiated by decision makers, it opens up possibilities to discuss other related aspects of a problem (11). Participation in that discussion will enable increasing understanding of statistical methodologies and application. When decision-makers make an effort to understand the usefulness of statistics, they can learn to better apply them in practice. This same inquiry remains necessary for statistics derived from validated and standardized measurements. By repeated inquiry into the construction of statistics, decision-makers gain a broader common understanding of statistics and their applicability. This thesis has provided a method to aid such an inquiry in the form of the FIAT-Health 2.0.

A method to improve the interpretation of statistics on health and healthcare

The FIAT-Health described in chapter 3 and 4 contributes to the systematic appraisal of characteristics relevant to the interpretation of statistics. Those with limited technical knowledge of statistical interpretation have few methods for the interpretation of a published statistic (19). The FIAT-Health is the first evidence-based instrument for the interpretation of publicly reported statistics. It fills the critical gap between the expert and members of the public whose decisions depend on the correctness of a statistic and its reporting (e.g., policymakers and advisors, journalists, managers, patients). The tool is not intended to make a 'hard judgement' by generating a use or not-use result, but helps the user to put better estimate the value of the statistic and to put it into a realistic perspective. Interpretation of statistics is human work. Any policy or practice decision requires balancing different aspects of which the statistical information is only one. By using the tool, one reflects on the characteristics relevant to the interpretation of statistics and its usability in different contexts.

Researchers, science communicators/journalists, knowledge integration specialists and policy advisors may use the FIAT-Health to assure all characteristics relevant to the interpretation of statistics are reported on in societal publications. This includes the funder of part I of this thesis, the National Institute for Public Health and the Environment (RIVM). By transparently reporting the background of statistics, political arguments meant to discredit the validity of the statistic can be anticipated. Policy debate may thus focus on the issues at hand, rather than the methodological restrictions of the evidence used. If they apply the tool to improve their reporting, communication between those creating the statistic i.e. the research institutions and the users of the statistic i.e. decision-makers can be enhanced (12).

There remains a question on how the FIAT-Health can be implemented in practice. The FIAT-Health was used by knowledge integration specialists at the RIVM to check the statistical reporting in the public online knowledge platform "The State of Health and Health Care" ("de Staat van Volksgezondheid en Zorg"). Although the FIAT-Health was useful in detecting areas of improvement in publicly reported statistics, it was not further implemented in the reporting practices of the institution. Routine embedding of the FIAT-Health in the work of statistical reporting requires sufficient incentives (13). Unlike scientific checklists, such as the EQUATOR guidelines (14), that can be referenced in scientific papers or required along with the submission to many scientific journals, there are no such direct incentives instituted to show accountability in responsible societal reporting through the help of an appraisal instrument. Implementing the FIAT-Health in routine reporting of statistics thus requires active efforts by a thought leader (13). This thought leader needs to dedicate continuous effort in engaging knowledge integration specialists in using the tool in their routines. The thought leader needs to develop knowledge integration specialists' skills for applying the tool in practice (e.g., organizing workshops for skill development). And she or he should implement structures that allow for the sustainable implementation of the tool in practice (e.g., accessibility of the tool in the work environment). Other international agencies that report statistics on a regular basis can implement the FIAT-Health in a similar fashion.

The FIAT-Health 2.0 may additionally prove useful in the education of the interpretation and responsible reporting of statistics, for instance in public health and medicine studies. Outside the health sector, integration of the tool into courses aimed at improving critical reflection on science could aid in the education of health statistics to students in political science or public administration. By applying the tool, students can learn to critically appraise publicly reported statistics and gain insight into what information is important to report when writing about a statistic. Consequently, they will naturally develop their ability to interpret statistics derived from public sources.

Moreover, the FIAT-Health can aid national news agencies such as the Dutch NOS, the British BBC, or the American NPO, as well as research institutes that publish press releases. The FIAT-Health can form a model for future 'fact-checking' and might be used to structure such activities. Journalist activities on fact checking are increasing (15). Websites such as *http://nieuwscheckers.nl/* in the Netherlands (16) and, *https://fullfact.org/* in the UK (17) focus fully on fact-checking publicly reported statements. Simultaneously, national and local news outlets have been instituting their own fact checking departments, such as the Volkskrant, BBC, and NPO. Although these separate departments increase attention to the responsibility of journalists to verify public statements, critical examination was always the core part of their profession. As a participant at a science communication workshop once said: "In the past, fact checking was always around, it was called 'journalism'!"

RESPONSIBLE REPORTING OF HEALTH SERVICES RESEARCH

The findings of part II help to facilitate critical reflection on the current state of research and encourage the advancement of responsible scientific and societal reporting of messages and conclusions in HSR.

Responsible scientific reporting of HSR

A responsible scientific publication is generally understood to be complete, accurate and transparent (18). Rules to achieve this level of reporting are well established and presented in scientific publication checklists. These checklists provide straightforward directions for the structure of a scientific publication, the inclusion of study information in the methodology section, description of the results section, description of methodological limitations, and reference to related literature. Moreover, there is an abundance of literature describing the restrictions and allowances for inferencing causality and generalizability from the research findings. These subjects necessary for a responsible scientific publication are all included in the definition of QRPs presented in chapter 5 of this thesis.

Our findings indicate that, like in biomedical and psychological disciplines, there is room for improvement on each of these subjects in scientific HSR reporting. However, two subjects related to responsible scientific reporting of HSR included in our definition of QRPs require specific reflection: the mention of contradictory evidence and the necessity to include implications and recommendations for policy and practice.

The absence of contradictory evidence in a scientific publication, or a description thereof, is part of the definition of QRPs in the reporting of messages and conclusions in HSR. In most of the reviewed scientific HSR publications, contradictory evidence was not reported. From the discussions held with the leaders of HSR institutions, no strong consensus resulted on the inclusion of contradictory evidence in scientific HSR publication. HSR is usually based on data derived from such real-life settings. In many cases, an author may argue that their context dependent studies cannot be compared directly with any other available evidence, and she or he cannot report contradictory evidence because it simply does not exist. Consequently, they omit any mention of it.

However, neglecting to mention contradictory evidence hinders interpretation of findings in the full context of evidence. If an author searched for contradictory studies and did not identify any, readers of the publication would not have any clues on whether the author included all relevant evidence. The assumption that if no contradictory evidence was mentioned, it is not available, is too implicit, and does not hold up to current standards of transparency in scientific literature. Therefore, responsible reporting in HSR includes a statement on contradictory evidence, even if it does not exist.

The definition of QRPs in the reporting of messages and conclusions additionally includes the reporting of implications and recommendations for policy and practice. The large amount of inadequately reported implications and recommendations are in direct contrast with the HSR community's efforts to strengthen the interaction between research and society (19-21). Despite the presence of items on implications and recommendations in commonly used publication checklists, the HSR community still disagrees on whether they should always be included in scientific publications. Moreover, in the focus groups described in chapter 7, it was discussed that researchers do not always recognize the importance of a recommendation section. Critics to including implication and recommendation sections in scientific

HSR publications argue that, like any other scientific discipline, HSR needs to shed light on complex systems and re-examine the status quo (22). Results from research that challenge normative thinking or evaluate complex systems are considered unsuitable by critics for immediate implementation (22). Moreover, the application of biomedical evidence to clinical practice differs from HSR. When evidence is applied to a political or professional environment, there are more non-scientific values to consider. Critics argue that decision-makers would be in a better position to weigh those values than the researcher could (22). Critics further state that the word limit of a publication does not allow an author to write a comprehensive overview of possible implementation of their results (23).

However, in this thesis, it is argued that recommendations and implications for policy and practice form a necessary component of a responsible scientific publication. HSR has a particular connection to policy and practice, as it aims to provide knowledge for the direct application of results to policy and practice (24). Assuring applicability to policy and practice can therefore be considered an inherent part of the responsibility of a health services researcher. A solid description of implications and recommendations offers a foundation for decision-making. When the advice provided in peer-reviewed articles is too abstract to implement directly, decision makers may not have any vision on how to apply the evidence to practice, both immediately and in the long term (25). Scientific publications providing managerial knowledge (i.e., program evaluations or interventions) are more easily understood than complex theoretical knowledge. Decision makers often lack the time to fully appraise evidence to their specific situation (26). Therefore, they are in a worse position to weigh scientific evidence against other values. A health services researcher who does have in-depth insight pertaining to the context of their research might not necessarily be responsible for stating ‘what needs to be done’, but does have the responsibility to transparently describe possibilities for practical application. Moreover, as described in chapter 6, neglecting to provide implications and recommendation for policy and practice in scientific publications opens the possibility for framing new or arbitrary recommendations in societal literature that might not be adequately justified by the findings. These possibly misinterpreted messages may become part of a greater political and societal debate (27-29) by indirectly influencing the decisions of decision-makers (30). It is the responsibility of a health services researcher to anticipate societal dissemination and offer a sound basis for policy and practice implementation in a scientific publication. Hence, the inclusion of implications and recommendations for policy and practice are a necessary part of a responsibly reported scientific HSR publication.

The overview of possible QRPs in the reporting of messages and conclusions in HSR developed in chapter 5 of this thesis includes these items. Together with existing publication checklists, the overview of QRPs developed in this thesis serves as a comprehensive guide for the contents of a responsible HSR publication.

Responsible societal reporting of HSR

As opposed to scientific publications, a societal publication offers more leeway to report scientific results. There is little consensus as to what constitutes responsible societal publication of HSR. The findings discussed in part I of this thesis indicated that a responsible societal publication ideally includes all characteristics of a scientific study that facilitate an adequate interpretation of the presented findings. Chapter 3 of this thesis provides these characteristics for statistics specifically. Nevertheless, most societal publications are limited in space and cannot include all this information. Any tweet, LinkedIn post, news article or factsheet needs to present information concisely. In those cases, a reference to the original scientific publication should be a minimum requirement for a responsible societal publication.

Additionally, a responsible societal publication accurately reflects the findings and discussion presented in the scientific publication. In particular, the unsupported introduction of causality or exaggerated conclusions is considered questionable or even misleading. Irresponsible interpretation of causality and exaggerated conclusions can inadvertently misguide evidence-based discussions. If these unfounded messages give reason to change policy or practice, they might leave a negative impact on the development of health and healthcare. Our research findings presented in chapter 6 indicate that beyond causality and exaggeration, more types of inconsistencies should not be overlooked. Is a societal publication questionable if it reports conclusions incompletely, provides a reinterpretation of findings and conclusions, or endorses recommendations that were not included in the corresponding scientific publication? Few changes can be made in a societal publication if the author intends to reflect the findings of a scientific publication accurately.

However, exact consistency does not always help write a comprehensible societal publication. Descriptions in scientific publications may be too technical to understand or there might be too little space to include all important limitations of a study. Moreover, messages and conclusions as reported in scientific publications could simply be too dull to attract attention of decision-makers necessitating simplification or rephrasing of messages and conclusions.

Then how can an interesting yet responsible societal publication be produced? Health services researchers are often not educated to write societal publications and are not always proficient in simplifying their scientific messages in a responsible manner (31, 32). In universities, researchers are pressured to maintain a high level of scientific publications and balance this primary task with education, teaching, and writing research grants. Moreover, researchers may lack awareness of outlets that are best suited to disseminate their message. They might not consider writing societal publications to be part of their profession and lack time to dedicate effort to improving these messages. To assure translation of research findings to society, resources need to be available and incentives need to be implemented for researchers to increase their ability and commitment to these activities. Employing staff with experience in societal dissemination dedicated to coordinating research and practice interaction could alleviate responsibility of researchers. Furthermore, science communicators specialize in the translation of research findings. Nevertheless, a risk arises that in simplification of research findings the message loses its consistency with the scientific publication. Researchers and science communicators should therefore collaborate. Science communicators would be able to simplify findings to make them interesting to decision-makers, whereas researchers would guard consistency with the scientific publication.

A responsible societal HSR publication should thus always include a reference to the source, and its content should involve a balance between readability and consistency with the underlying scientific publication. To achieve this, increased interaction and understanding between the researchers and professional science communicators is necessary.

Responsibility of Health Services Research institutions

The process of scientific publication is highly structured to support responsible scientific reporting, (14). National funders increasingly require open access publication, data sharing, and strategies for creating societal impact (20, 33, 34). Scientific journals have implemented various measures to stimulate transparency in scientific publications. This includes adherence to publication guidelines, open-access publication, statements on data sharing, and, in some cases, post-publication peer-review and committing to publication based on study protocol. Funders and scientific journals need to continue strengthening these measures (35).

However, because HSR institutions shape the main environment where a scientific manuscript is written, they have the largest opportunity to support researchers in writing responsible publications. Dutch HSR institutions have a variety of support structures for their researchers. HSR institutions internationally can benefit from the lessons learned in the Dutch context. A wide heterogeneity in (post-) graduate education and requirements exists. For instance, some institutions enforce education either through mandatory courses or a set amount of credits whereas other institutions have not set fixed educational requirements. Moreover, across HSR institutions, the minimum number of scientific publications necessary to complete a PhD thesis ranges from one-to-four accepted peer-reviewed publications. The causes of these differences lay partially in the demands of their respective Graduate schools, but also in institutional expectations. Moreover, HSR institutions are embedded in different disciplinary faculties (e.g., medicine, sociology, administrative sciences, economic and management faculties). A biomedical foundation would produce different expectations (perhaps a stronger ambition to publish in journals with a high impact factor) than a sociological foundation. Although these differences in norms are not necessarily detrimental to the field at large, a strong focus on scientific publication leaves less room for consideration of possible practical implications of research findings. Moreover, a higher pressure to publish reduces time available for specific education in responsible societal or scientific publication. In chapter 7, it was found that specific training in writing messages and conclusions was related to a lower number of QRPs. Unfortunately, these training opportunities are not equally accessible at all institutions. Some institutions only provide basic writing training while others offer more rigorous writing courses. HSR institutions do align when it comes to the organization of supervision. Currently, according to university regulation in the Netherlands, at least two supervisors are strongly involved in the writing process of a scientific manuscript.

Differences in support structures do not only apply to at the PhD educational level. Quality assurance at the start of a research project strongly diverge amongst institutions (i.e. the writing of a research proposal, and the final part of a research project, including societal dissemination of findings). Institutions have implemented several peer-feedback structures. Peer-feedback is stimulated at different stages of the research process, including the design, execution and analyses of a project, the interpretation of findings, and the writing of a manuscript. Particularly in the final stages of a manuscript, peer-feedback helps to correct QRPs overlooked by co-authors. Some institutions provide multiple opportunities for peer-feedback at each stage of the research project. Other institutions maintain a researcher's freedom to seek out these opportunities when she or he deems support necessary. Peer-feedback is often a welcome addition to an HSR institution, as is indicated by voluntarily peer-feedback groups formed by PhD students at several institutions. However, discussions in the focus groups described in chapter 7 indicate that when peer-feedback groups are voluntary, researchers often avoid participation. Moreover, feedback-structures at the HSR institution (in their current form) were not generally associated with a lower number of QRPs in scientific publications. A large concern for researchers is their time investment in participating in these peer-feedback sessions. They lack the conviction that participation will benefit their work, or exhibit reluctance to receive negative feedback (eventually resulting in extra work) from colleagues. Consequently, to successfully implement peer-feedback structures, the benefits of participating in peer-review groups and receiving feedback at all levels of seniority needs to be made clearer to researchers, and research institutions need to foster a positive feedback culture in these groups.

With the absence of (successful) peer-feedback opportunities at some institutions in the Netherlands, available support for post-doc and senior researchers is especially limited. They can work independently on individual research projects without much oversight. Mid-career researchers are under pressure to obtain grants and conduct individual research projects while maintaining a high level of scientific publication. The necessity to become a “jack of all trades” decreases the freedom of researchers to ensure the quality of work in each of these responsibilities. Developing talent in writing grant proposals differs from developing analyses techniques or project management, with each of these responsibilities requiring extensive study and practice. Consequently, support at each stage of the research process is needed to assure responsible scientific publication. HSR institutions need to inquire whether their more senior researchers have the time and resources to pursue opportunities for feedback from their peers.

RECOMMENDATIONS FOR POLICY AND PRACTICE

The recommendations below are described in the order that scientific research is produced , starting with HSR funders, leaders of HSR institutions, scientific journal editors, individual researchers, science communicators/ journalists, and ending with decision-makers: the users of the research findings (i.e. policy makers, healthcare professionals, the general public and patients).

HSR funders

To stimulate transparent and accountable scientific reporting of their research projects, funders should require measures that promote transparency and accountability in publication. These measures include open access publication, peer-review of study protocols, protocol registration, and public availability of collected data. In particular, national funding agencies have increasingly set requirements to stimulate transparent research. Given the results of this thesis, it is recommended additionally that non-national funders to include these demands in their funding proposals as well.

The Netherlands Organization for Health Research and Development (ZonMw) is the national Dutch funder of research on health and healthcare. To stimulate better interpretation and practical relevance of research, they and other funders should dedicate a separate budget to stakeholder involvement.

To improve recommendations and implications for policy and practice in scientific publications, a strong connection between the research and practice is needed. This connection can be established through stakeholder involvement in research projects. In particular, stakeholder contributions may help researchers develop ideas for connecting their findings to policy and practice. It further strengthens the perceived credibility of research results, assists in the development of quality research tools, increases researcher access to potential study participants, facilitates multi-perspective interpretation and analysis, and supports strategic dissemination of key findings (36). HSR project leaders may sacrifice stakeholder involvement because of limited time and resources available in a research project. Moreover, the risk arises that stakeholders might influence the research through political interests. To responsibly involve stakeholders in the research process, their contribution to the research should be thoroughly planned. Funding can be set aside to hold stakeholder meetings involving relevant patient and professional organizations to achieve this. Moreover, increased focus on societal impact requires a professionalization of societal reporting and dissemination of research findings. A specific budget and expertise in responsible dissemination of research findings should be demanded in project proposals to allow for these activities.

To continue improving responsible reporting of HSR and other scientific disciplines, funders need to finance research on research. Increasingly, national funders have invested in these efforts and should continue to do so.

Leaders of HSR institutions

HSR institutions in the Netherlands and internationally should create opportunities for interaction between health services researchers to stimulate a supportive research climate across the field. The HSR community in the Netherlands can collaborate on sharing courses on responsible reporting of HSR with specific attention paid to the interpretation and reporting of messages and conclusions.

HSR leaders should implement peer-feedback sessions at different stages of writing a manuscript to stimulate discussions on design, analyses, and particularly interpretation of research findings. Making such regular feedback sessions mandatory parts of the institute's quality system can guarantee a stronger embedding. Several HSR institutes in the Netherlands have implemented strong feedback structures that can serve as an example for other HSR institutions. When implementing such structures, HSR institutes should assure the time required from their researchers benefits their work. Aside from being beneficial to junior researchers, such peer-feedback structures could serve to support senior researchers who conduct independent research projects. The definition of QRPs in the reporting of messages in HSR may provide helpful in providing feedback on messages and conclusions.

Moreover, to adequately plan stakeholder involvement in research projects, a stakeholder contribution plan and suitable budget should be included in research proposals. If funders reward these efforts, health services researchers can invest adequate time in adequate stakeholder collaboration.

In research training, HSR institutions may use the FIAT-Health to educate students in the interpretation of statistics. Students can also be taught how to approach public reporting of research findings. For example, teachers may develop an assignment that makes students assess a publicly reported statistic. Their assessment can be compared to an assessment through the FIAT-Health to reveal the often-overlooked aspects of a publicly reported statistic.

Finally, HSR institutes should professionalize their efforts to create societal impact. Researchers may not have the ability or time to learn responsible societal publication practices. Institutions should therefore employ staff dedicated to societal dissemination. This societal dissemination specialist could develop skills and strategies for the responsible societal publication of HSR findings from within the research institution.

Scientific journal editors

Scientific journals editors can further contribute to more responsible scientific publication by increasing their focus on open science measures, demanding data transparency, requiring pre-publication protocols, or demanding registered protocols and research proposals. For further research dissemination, they can stimulate open access publications. Publication based on research proposals may diminish pressure to exaggerate conclusions. Moreover, editors and reviewers should pay more attention to the content of the implications and recommendations sections in a scientific HSR publication.

Journals frequently tweet or provide a press release based on a scientific publication. Communication with the authors of the scientific publication would improve responsible messaging. Moreover, journals should manage press releases to assure that aspects most important to the interpretation of research findings are reported.

Individual researchers

Researchers themselves can use the assessment scheme of QRPs presented in this thesis when reporting messages and conclusions. Researchers are advised to seek additional feedback from their co-authors and peers if they deem it necessary, or organize feedback exchanges with their colleagues if no such structure is offered at their respective institutions. Although going through the feedback process may be time consuming, the benefits of a feedback structures are likely to outweigh costs.

When writing societal publications on their research, researchers should include characteristics that are necessary for accurate interpretation of evidence (e.g. FIAT-Health items). Researchers are in the best position to guard the consistency and accuracy of any societal publication with respect to the original scientific publication. However, researchers should also trust science communicators or journalists to simplify their research findings in a responsible manner. Strong collaboration between these parties may improve responsible dissemination of research findings.

Science communicators/journalists and knowledge integration specialists

Science communicators and journalists are advised to use the FIAT-Health in fact checking efforts when writing a societal publication related to statistics. Platforms dedicated to fact checking can use the FIAT-Health as an instrument to better understand the construction of a statistic.

Science communicators and journalists should be aware that a single study is often not substantive enough to provide adequate content for a news item or an interesting press release. Nevertheless, results from scientific publications are often presented that way. Rather than describe the results of one study, it is more worthwhile to address the overarching context of the study and what new results may add to that context. Science communicators and journalists should interact with the authors of the scientific publication to explore and identify the relevant context of their study.

In the Netherlands, the RIVM plays a pivotal role in the public dissemination of policy relevant statistics on health and healthcare. Knowledge integration specialists at the RIVM control multiple websites where many available statistics on the health of the Dutch population can be found. In an attempt to create clarity in the abundance of available statistics, the Dutch ministry has commissioned RIVM to develop and maintain one website that functions as the main source of statistics on health and healthcare (https://www.staatvenz.nl). The RIVM may use the FIAT-Health to assure the responsible societal reporting of statistics on their platforms. They might integrate the FIAT-Health in their continuous efforts to improve information dissemination to decision-makers. Knowledge institutes internationally are encouraged to do the same.

Policy makers and healthcare professionals

Policy makers and healthcare professionals should be aware that statistics and scientific evidence are developed for a certain purpose and specific context. When using HSR evidence, they should appraise whether the evidence is transferable to the situation in which they intend to use it. When communicating about statistics, policy makers and healthcare professionals should assure that they have adequate knowledge of the characteristics most important for their interpretation. They might apply the FIAT-Health for statistics on health and healthcare, but also other fields such as educational, environmental and economic disciplines.

Public and patients

The public and patients should remain critical of statistics and HSR evidence reported in societal publications. They are advised to question who was responsible for constructing the statistic, as well as the purpose of the statistic itself. When communicating with a care professional, people may question the statistical information they receive and discuss the value of these statistics in relation to their specific situation. The FIAT-Health could provide a strong foundation on which such a discussion could begin. People could generally apply the FIAT-Health to statistics they read online or hear about in the news to improve their own understanding of statistics.

FUTURE RESEARCH

Based on this thesis, several recommendations for future research can be formulated.

To provide recommendations for the integration of statistics in policy debate internationally, more research is needed to examine how statistics are currently used in policy and practice. Research is also needed to determine novel methods for facilitating responsible use of such statistics. Chapter 2 focused on the Dutch context; therefore, more knowledge is needed on the role of statistics in other countries' policy debates. Chapter 3 and 4 suggest research should be conducted on how to best integrate critical assessment of statistics and other HSR evidence in healthcare policy and practice.

This thesis further revealed a need for more research on how to best aid decision-makers with best interpreting research findings. Findings suggested a particular need for research on difficulties experienced when interpreting statistics on health and healthcare and other HSR evidence. Too little is known about the actual interpretation of research findings as opposed to the use of evidence in policy and practice. Research in this direction will provide insight in how HSR evidence can responsibly be integrated in society. Moreover, better knowledge on interpretation of research can stimulate guideline development for reporting implications and recommendations for policy and practice in scientific publications.

More research is required on the interplay between the reporting in scientific publications and societal publications. Chapter 6 shows that QRPs in scientific publications are often replicated in societal publications. When studying the quality of societal publications, these replicated QRPs are currently overlooked. To gain accurate insight into the responsible communication of scientific findings in societal publications, future studies both in and outside the field of HSR should integrate the replication of QRPs from scientific publications.

Additionally, chapter 7 provided an exploration of factors related to the reporting of HSR in scientific publications. To gain stronger insight into the contribution of these factors to scientific reporting, this research should be extended to a larger selection of researchers, and a more in-depth inquiry should be conducted on the identified factors. Because many factors affect scientific reporting simultaneously, studies should include an analysis of multiple factors, rather than focus on one factor of influence as is currently the case. In addition, further research should focus on the factors that contribute to more responsible societal publications as well.

Recent studies emphasize the decisive role of supervision in the conduct of scientific research (37). Supervision may play a significant role in the development of reporting skills in both junior and senior researchers. Consequently, the role of supervisors and how they can aid the responsible reporting of scientific research should be investigated. Further methods at the institutional level to improve researchers' competency in writing scientific and societal publications should be studied. A study examining various support structures across the field of HSR would provide better knowledge on how institutions can best aid their researchers in both scientific and societal reporting

CONCLUSION

By conducting the research in this thesis in collaboration with the HSR community, a debate was started on the responsible interpretation of statistics on health and healthcare, as well as the reporting of HSR in scientific and societal publications. The findings of this thesis provide a critical reflection on the use of statistics and current state of research. This thesis encourages HSR funders, leaders of HSR institutions, scientific journal editors, individual researchers, science communicators/journalists, and ending with decision-makers: the users of the research findings (i.e. policy makers, healthcare professionals, the general public and patients) to advance interpretation of statistics on health and healthcare, and responsible scientific and societal reporting of messages and conclusions in HSR.

To conclude, this thesis identified characteristics necessary to interpret publicly reported statistics on Health and Healthcare, and provided a method for individual researchers, science communicators/journalists, and decision-makers to appraise these characteristics in the form of the FIAT-Health. If individual researchers, science communicators/ journalists apply the tool to improve their public reporting practices, and decision-makers use the tool to better interpret statistics, the responsible use of statistics can be enhanced. Questionable Research Practices (QRPs) in the reporting of messages and conclusions occur frequently in peer-reviewed international scientific HSR publications. Particularly the reporting of contradictory evidence, and recommendations and implications for policy and practice require improvement. In addition to frequent occurrences of inconsistencies between scientific and societal publications, QRPs in the reporting of messages and conclusions are often replicated in societal publications. To address the current state of scientific reporting in HSR, supportive measures on specific writing training, pressure to create societal attention and co-author conflict of interests should be implemented by HSR funders, leaders of HSR institutions and scientific journal editors.

The question mark from the title of this thesis “Open to Interpretation?” will remain. However, the insights offered in this thesis can contribute to a responsible (HSR) research practice and improved interpretation of statistics and research findings.

REFERENCES

1. de Groot AD. Methodologie. Grondslagen van onderzoek en denken in de gedragswetenschappen. Assen: Van Gorcum; 1994. p. 130.
2. Jagosh J, Macaulay AC, Pluye P, Salsberg J, Bush PL, Henderson J, et al. Uncovering the benefits of participatory research: implications of a realist review for health research and practice. Milbank Q. 2012;90(2):311-46.
3. Daviter F. The political use of knowledge in the policy process. Policy Sciences. 2015;48(4):491-505.
4. Weiss CH. Knowledge Creep and Decision Accretion. Knowledge. 1980;1(3):381-404.
5. Espeland WN, Stevens ML. A Sociology of Quantification. Arch Eur Sociol. 2008;49(3):397-432.
6. Porter ME. What Is Value in Health Care? New England Journal of Medicine. 2010;363(26):2477-81.
7. Groenewoud AS, Westert GP, Kremer JAM. Value based competition in health care's ethical drawbacks and the need for a values-driven approach. BMC Health Services Research. 2019;19(1):256.
8. Campbell SM, Braspenning J, Hutchinson A, Marshall M. Research methods used in developing and applying quality indicators in primary care. Quality and Safety in Health Care. 2002;11(4):358-64.
9. Moes FB, Houwaart ES, Delnoij DMJ, Horstman K. "Strangers in the ER": Quality indicators and third party interference in Dutch emergency care. Journal of evaluation in clinical practice. 2019;25(3):390-7.
10. Anema HA, Kievit J, Fischer C, Steyerberg EW, Klazinga NS. Influences of hospital information systems, indicator data collection and computation on reported Dutch hospital performance indicator scores. BMC Health Services Research. 2013;13(1):212.
11. Essen A, Oborn E. The performativity of numbers in illness management: The case of Swedish Rheumatology. Soc Sci Med. 2017;184:134-43.
12. Spiegelhalter D. Trust in numbers. Journal of the Royal Statistical Society Series A: Statistics in Society. 2017;180(4):948-65.
13. May C, Finch T. Implementing, Embedding, and Integrating Practices: An Outline of Normalization Process Theory. Sociology. 2009;43(3):535-54.
14. Altman DG, Simera I, Hoey J, Moher D, Schulz K. EQUATOR: reporting guidelines for health research. Lancet. 2008;371.
15. Stencel M. Number of fact-checking outlets surges to 188 in more than 60 countries: Dukes Reporters Lab; 2019 [Available from: https://reporterslab.org/number-of-fact-checking-outlets-surges-to-188-in-more-than-60-countries/.
16. De Factcheckers van de Universiteit van Leiden. Nieuwscheckers 2019 [Available from: http://nieuwscheckers.nl/.
17. Full Fact. Full Fact fights bad information. 2019 [Available from: https://fullfact.org/.
18. Altman DG, Simera I. Responsible reporting of health research studies: transparent, complete, accurate and timely. Journal of Antimicrobial Chemotherapy. 2009;65(1):1-3.
19. Dinsmore A, Allen L, Dolby K. Alternative Perspectives on Impact: The Potential of ALMs and Altmetrics to Inform Funders about Research Impact. PLOS Biology. 2014;12(11):e1002003.

20. Council for Medical Sciences. The societal impact of applied health research: Towards a quality assessment system. Amsterdam: Royal Netherlands Academy of Arts and Sciences; 2002.
21. Canadian Institutes of Health Research. Developing a CIHR Framework to Measure The Impact of Health Research. Ottawa; 2005.
22. Lewis S. How has health services research made a difference? Healthcare policy = Politiques de sante. 2011;6(Spec Issue):74-9.
23. Cummings P. Policy recommendations in the discussion section of a research article. Injury Prevention. 2007;13(1):4.
24. Colby DC, Baker LC. Health services research in 2020: an assessment of the field's workforce needs. Health Serv Res. 2009;44(6):2193-7.
25. de Haas B, van der Kwaak A. Exploring linkages between research, policy and practice in the Netherlands: perspectives on sexual and reproductive health and rights knowledge flows. Health Res Policy Sy. 2017;15(1):40.
26. Innvaer S, Vist G, Trommald M, Oxman A. Health policy-makers' perceptions of their use of evidence: a systematic review. J Health Services Res Policy. 2002;7.
27. Lavis JN, Ross SE, Hurley JE, Hohenadel JM, Stoddart GL, Woodward CA, et al. Examining the role of health services research in public policymaking. Milbank Q. 2002;80(1):125-54.
28. Kuruvilla S, Mays N, Walt G. Describing the impact of health services and policy research. J Health Serv Res Policy. 2007;12 Suppl 1:S1-23-31.
29. Gerrits RG, van den Berg MJ, Klazinga NS, Kringos DS. Statistics in Dutch policy debates on health and healthcare. Health Res Policy Syst. 2019;17(1):55.
30. Burstein P. The impact of public opinion on public policy: A review and an agenda. Polit Res Quart. 2003;56(1):29-40.
31. Dudo A. Scientists, the Media, and the Public Communication of Science. Sociology Compass. 2015;9(9):761-75.
32. Yuan S, Besley JC, Dudo A. A comparison between scientists' and communication scholars' views about scientists' public engagement activities. Public Understanding of Science. 2019;28(1):101-18.
33. Nederlandse Organisatie voor Wetenschappelijk Onderzoek. Belangrijkste resultaten van de NWO/ZonMw-consultatiebijeenkomst over Plan S. 2019.
34. Zerhouni EA. Information access. NIH public access policy. Science (New York, NY). 2004;306(5703):1895-.
35. Bouter LM. Fostering responsible research practices is a shared responsibility of multiple stakeholders. J Clin Epidemiol. 2018;96:143-6.
36. Hinchcliff R, Greenfield D, Braithwaite J. Is it worth engaging in multi-stakeholder health services research collaborations? Reflections on key benefits, challenges and enabling mechanisms. Int J Qual Health Care. 2014;26(2):124-8.
37. Haven T, Tijdink J, Pasman HR. et al. Res Integr Peer Rev. 2019;4:25.

Appendix

SUMMARY

Terms like "fact free politics", "science as an opinion", and "alternative facts" signal an uneasiness in society regarding the credibility of science's role in politics and policy making. When research findings are disseminated outside the scientific community, they are often reinterpreted and rephrased. Although useful in some cases, this process also increases risk for misinterpretation. That is, misreported research findings can inadvertently mislead policy makers, professionals, and the public, and misdirect the advancement of professional practice and policy. Directly or indirectly, research has a lasting influence on society. To foster the responsible interpretation and reporting of research findings to decision makers outside of the research community, a continuous reflection of the execution and translation of research to practice is required.

This thesis focuses on Health Services Research (HSR), which often meets broad societal interest by addressing topics such as co-payments, evaluation of quality improvement efforts, cost-effectiveness of medications, patient empowerment, compliance with therapy, and effects of policies. To facilitate increased support for responsible research practices in HSR, it is vitally important to explore how researchers report findings in scientific and societal literature, as well as how users interpret data and statistics on health and healthcare.

Thus, this thesis addresses the following aims:

1) Propose a method to improve the interpretation of publicly reported statistics on health and healthcare, and

2) Provide insight into the scientific and public reporting of Health Services Research.

Part I of this thesis (i.e., chapters 1 to 4) addresses the interpretation of statistics on health and healthcare.

In **chapter 2,** the use of statistics in the policy debate is evaluated via an analysis of Dutch policy documents published between 2014 and 2016 on eight different health and healthcare-related topics. We found that sources of the statistics used in policy debates were primarily government-funded. We identified two distinct functions of statistics in these documents (i.e., managerial and rhetorical). Statistics that were used for planning, monitoring or evaluation of policy were considered to be used managerially. Statistics with the managerial function of policy evaluation were mostly the result of routine or standardized data collection. When a statistic was used for agenda setting or to convince others of the importance of a topic, we called this rhetorical use. When statistics had a managerial function, policy makers mirrored terms used by the scientific community in their discussions. Policy makers used terms such as statistical significance and case-mix. When statistics were used for a rhetorical function, this was not the case.

Next, this thesis discusses how statistics used for rhetorical purposes settle in our common understanding and may change our values and priorities. Relying only on statistics to determine the relative importance of various social problems limits consideration of non-measurable values, such that several of these values, while critical to the political debate, are often forgotten. A critical reflection on statistics that have a rhetorical function, is at least as important as a reflection on statistics used with a managerial purpose. If statistics are not carefully evaluated and discussed, they might not be applied correctly and misdirect the progress of healthcare rather than help it forward.

Chapter 3 describes the development of a tool to facilitate better interpretation of statistics on health and healthcare. Identified characteristics relevant to the interpretation of statistics were their origin, credibility, expression (i.e. percentage or absolute number), subject matter, population and geographical focus, time period, and underlying data collection methods. These characteristics were translated into questions that were used to construct the Figure Interpretation Assessment Tool – Health (FIAT-Health) 1.0. This tool contains a set of 13 dichotomous and 4-point Likert scale questions along with two final assessment statements. The testing and improvement of the FIAT-Health 1.0 is described in **chapter 4**. Results from our study indicated that questions in the FIAT-Health 1.0 were comparable to the type of questions used in expert assessments. Potential users reported that the format and language needed improvement. Therefore, the tool was refined and transformed from a quantitative scoring instrument into an online qualitative appraisal tool: the FIAT-Health 2.0. The FIAT-Health 2.0 consists of 13 factual questions, questions regarding the agreement between the primary research publication and the public report, and two open-ended assessment questions. The FIAT-Health 2.0 can help policymakers, communication officers, and researchers systematically interpret publicly reported figures on health(care).

Before any statistic can be used responsibly in policy or practice, decision makers should examine and discuss it critically. If decision-makers question statistics and how they are constructed, they may also think of other non-measurable aspects of the issue at hand. Such a discussion will enable increasing understanding of statistical methodologies and application. When decision-makers make an effort to understand the usefulness of statistics, they can learn to better apply them in practice and gain a broader common understanding of their applicability. This inquiry is not only important for results from smaller studies, but is also necessary for statistics derived from validated and standardized measures.

This thesis proposes one method to aid such an inquiry in the form of the FIAT-Health 2.0. The tool is not intended to make a 'hard judgement' by generating a use or not-use result. Instead, it helps the user to correctly estimate the value of the statistic and put it into perspective. Interpretation of statistics is human work. Any policy or practice decision requires balancing different values of which the statistical information is only one. By using the tool, one reflects systematically on the characteristics relevant to the interpretation of statistics and their usability in different situations.

Part II of this thesis (i.e., chapters 5 to 7) addresses the responsible reporting of Health Services Research in scientific and societal publications.

Chapter 5 describes the occurrence and nature of Questionable Research Practices (QRPs) in the reporting of messages and conclusions in scientific HSR publications. Together with 13 participating HSR institutions in the Netherlands, we defined QRPs in the reporting of messages and conclusions in HSR and developed an assessment instrument covering 35 possible QRPs. Using this assessment instrument, we assessed a random sample of 116 HSR articles authored by researchers from participating HSR institutions published in international peer-reviewed scientific journals in 2016. A median of six QRPs per publication was identified. QRPs occurred most frequently in the reporting of implications for policy and practice, recommendations for policy and practice, contradictory evidence, study limitations, and conclusions based on the results and in the context of the literature. We identified no differences in the total number of QRPs in publications based on different methodological approaches, type of research, or study design.

A responsible scientific publication is complete, accurate and transparent. Two topics related to responsible scientific reporting of HSR included in our definition of QRPs require specific reflection: the mention of contradictory evidence and the necessity to include implications and recommendations for policy and practice.

Many of the reviewed scientific HSR publications did not present contradictory evidence when discussing results, in contrast with current standards of transparency in scientific literature. However, this thesis argues that responsible reporting in HSR should always include a statement on contradictory evidence, even if it does not exist.

Further, this thesis concludes that refraining from describing implications and recommendations for policy and practice in scientific publications opens the door for societal publications to create arbitrary recommendations that might not be adequately justified by the research findings. These possibly misinterpreted messages may become part of a wider debate and unjustly influence decision-makers. It is the responsibility of a health services researcher to anticipate societal dissemination and to provide a sound basis for policy and practice implementation in a scientific publication. Hence, the inclusion of implications and recommendations for policy and practice should be a necessary part of a responsibly reported scientific HSR publication.

Chapter 6 describes the analyses of replicated QRPs in the reporting of messages in societal HSR publications, along with inconsistencies with the original scientific HSR publications. We conducted a qualitative content analysis of societal publications derived from the scientific HSR publications assessed in chapter 5. Results indicated that most societal publications contained some inconsistencies or replicated QRPs derived from the corresponding scientific publications. These issues were most often characterized by inadequately reported conclusions, policy and practice recommendations, and titles. No substantial differences in number of QRPs emerged between societal publications produced by authors of the scientific publication, societal publications published on institute or funder pages, and societal publications written by a person who was not involved in the scientific publication.

As opposed to scientific publications, a societal publication offers more leeway when reporting scientific results. The findings discussed in part I of this thesis indicate that a responsible societal publication ideally includes all characteristics of a scientific study that facilitate an adequate interpretation of the presented findings. However, societal publications usually have limited word space, so not all characteristics can be mentioned. A reference to the original scientific publication should, however, be a minimum requirement for a responsible societal publication. Additionally, a responsible societal publication should accurately reflect the findings and discussion presented in the underlying scientific publication. It should be noted that firm consistency does not always help when writing a comprehensible societal publication.

Descriptions in scientific publications may be too technical to understand or there might be too little space to include all important limitations of a study. Further, messages and conclusions as reported in scientific publications could simply be too dull to attract attention of decision-makers, thus necessitating simplification or rephrasing of messages and conclusions.

Simplifying research findings is not without its risks. A message in an "easier-to-read" societal publication may lose its consistency with the scientific publication. Hence, contents of a responsible societal HSR publication should maintain a balance between readability and consistency with the underlying scientific publication. A responsible societal publication should thus be formed through increased interaction and understanding between the researchers and professional science communicators.

Finally, **chapter 7** describes factors related to QRPs in scientific publications. Findings suggested that pressure to create societal impact was associated with a higher number of QRPs in the reporting of messages and conclusions in HSR publications. Specific training in reporting messages and conclusions, as well as an awareness of co-author conflicts of interests, were related to fewer QRPs in HSR publications.

The factors identified in this chapter may assist the HSR community in supporting responsible scientific publication. Because HSR institutions shape the main environment where scientific manuscripts are written, they have the largest opportunity to support researchers in writing responsible publications. Support at each stage of the research process is needed to assure responsible scientific publication practices. Training opportunities are not equally accessible at all Dutch HSR institutions. Some institutions only provide basic writing training while others offer more rigorous writing courses. Institutions have also implemented several peer-feedback structures. Some institutions provide multiple opportunities for peer-feedback at each stage of the research project. Other institutions maintain a researcher's freedom to seek out these opportunities when she or he deems support necessary. However, our research indicates that when peer-feedback groups are voluntary, researchers often avoid participation. To implement successful feedback structures, the benefits of participating in peer-review groups and receiving feedback at all levels of seniority needs to be made clearer to researchers. Research institutions need to foster a positive feedback culture.

Conclusion

To conclude, this thesis identifies characteristics necessary to interpret publicly reported statistics on Health and Healthcare and proposes a tool to help individual researchers, science communicators/journalists, and decision-makers appraise these characteristics (i.e., FIAT-Health 2.0). If individual researchers and science communicators/journalists apply the tool to improve their public reporting practices, and if decision-makers use the tool to better interpret statistics, the responsible use of statistics can be enhanced. Questionable Research Practices (QRPs) in the reporting of messages and conclusions occur frequently in peer-reviewed international scientific HSR publications produced by Dutch institutions. Improvement is necessary in the reporting of contradictory evidence, and recommendations and implications for policy and practice. In addition to frequent occurrences of inconsistencies between scientific and societal publications, QRPs in the reporting of messages and conclusions are often replicated in societal publications. To address the current state of scientific reporting in HSR, HSR institutions should take measures to offer specific writing training in the reporting of messages and conclusions. Further, researchers should be attentive to possible conflicts of interests of co-authors. Finally, HSR funders, leaders of HSR institutions and scientific journal editors should strengthen policies that decrease pressure on researchers to create societal impact.

The question mark from the title of this thesis “Open to Interpretation?” will remain. However, the insights offered in this thesis can contribute to responsible (HSR) research practices and improved interpretation of statistics and research findings.

NEDERLANDSTALIGE SAMENVATTING

Uitdrukkingen zoals "feiten-vrije politiek", "wetenschap is maar een mening", en "alternatieve feiten", laten een onbehagen zien in de samenleving over de geloofwaardigheid van de rol van wetenschap bij politieke besluitvorming en beleidsvorming. Wanneer onderzoeksresultaten buiten de gemeenschap van wetenschappers worden verspreid, ontstaat een risico voor misinterpretatie. Tekortkomingen in onderzoeksonderzoeks rapportages kunnen beleidsmakers, professionals en het publiek misleiden, en praktijk en beleid in een onbedoelde richting sturen. Onderzoek heeft direct of indirect een blijvende invloed op de samenleving. Om een verantwoorde interpretatie en rapportage van onderzoeksresultaten aan beleidsmakers, professionals en het algemene publiek te bevorderen, is een voortdurende reflectie op de rapportage en vertaling van onderzoek naar de praktijk vereist.

In dit proefschrift staat gezondheidszorgonderzoek (GZO) centraal. De vraagstukken waar GZO zich op richt staan meestal volop in de maatschappelijke belangstelling en betreffen onderwerpen zoals eigen bijdragen in de zorg, evaluatie van kwaliteitsverbetering, kosteneffectiviteit van medicijnen, versterking van de positie van patiënten, therapietrouw en effecten van gezondheidszorgbeleid.

Door te bestuderen hoe gebruikers van wetenschappelijke data en cijfers over gezondheid en gezondheidszorg deze interpreteren, en hoe onderzoekers resultaten van gezondheidszorgonderzoek rapporteren, kunnen we een verantwoorde GZO-praktijk bevorderen.

Dit proefschrift heeft de volgende doelstellingen:

1) het ontwikkelen van een instrument voor de ondersteuning van een betere interpretatie van cijfers over gezondheid en gezondheidszorg, en

2) het verschaffen van inzicht in de wetenschappelijke en publieke rapportering van gezondheidszorgonderzoek.

Deel I van dit proefschrift, (hoofdstukken 1-4), richt zich op de interpretatie van cijfers over gezondheid en gezondheidszorg.

In **hoofdstuk 2** is het gebruik van cijfers in het publieke debat over gezondheidsbeleid onderzocht door een analyse uit te voeren van beleidsdocumenten te analyseren, die gepubliceerd zijn tussen 2014 en 2016 over acht verschillende gezondheid- en gezondheidszorg gerelateerde onderwerpen.

Het blijkt dat bronnen van de gebruikte cijfers in beleidsdiscussies voornamelijk publiek gefinancierd zijn. In de onderzochte documenten hebben we twee verschillende functies voor het gebruik van cijfers geïdentificeerd, namelijk een retorische functie en een management- gerelateerde functie. Wanneer een specifiek cijfer wordt gebruikt voor het overtuigen van de lezer of toehoorder van het belang van een onderwerp, noemen we dit retorisch gebruik. Cijfers die gebruikt worden voor de planning, monitoring of evaluatie van beleid vallen onder de managementfunctie. Bij de evaluatie van specifiek beleid waren de toegepaste cijfers meestal het resultaat van een routinematige of gestandaardiseerde dataverzameling. Wanneer gebruik van cijfers een management doel heeft, nemen beleidsmakers termen over uit het wetenschappelijke discours. Dit betreft bijvoorbeeld termen als "statistische significantie" en "case-mix". Wanneer cijfers retorisch gebruikt worden, was dit niet het geval.

In de discussie van dit proefschrift wordt gesteld dat wanneer cijfers worden gebruikt voor retorische doeleinden deze zich nestelen in onze gemeenschappelijke wijze van het begrijpen van gezondheid en gezondheidszorg en daarmee onze waarden en opvattingen over prioriteiten kunnen beïnvloeden. Echter, niet alles van waarde is meetbaar. Andere belangrijke waarden die relevant kunnen zijn voor het politieke debat worden daardoor uitgesloten. Een kritische reflectie op cijfers die retorisch gebruikt worden is minstens zo belangrijk als een reflectie op cijfers die gebruikt worden voor managementdoeleinden. Als cijfers niet zorgvuldig bediscussieerd worden, kan de bedoelde wetenschappelijke fundering worden aangetast.

De ontwikkeling van een methode om een betere interpretatie van cijfers over gezondheid en gezondheidszorg te vergemakkelijken, wordt beschreven in **hoofdstuk 3.** Voor de interpretatie van cijfers worden de volgende kenmerken relevant geacht: de herkomst van het cijfer, geloofwaardigheid, typering (bijv. percentage of absoluut getal), onderwerp, populatie en geografische gebied, tijdsperiode en onderliggende methoden van tellen en meten. Kenmerken werden omgezet in vragen die werden gebruikt om de "Figure Interpretation Assessment Tool – Health (FIAT-Health) 1.0 te ontwikkelen. Het instrument bevat een set van 13 dichotome en 4-punts Likert-schaalvragen en twee eindevaluatie- vragen. Het toetsen en evalueren van de FIAT-Health 1.0 wordt beschreven in **hoofdstuk 4.** Resultaten van deze studie geven aan dat vragen in de FIAT-Health vergelijkbaar waren met het type vragen dat werd gebruikt in beoordelingen door experts. Potentiële gebruikers gaven aan dat opzet en taal van de FIAT-Health 1.0 verbetering behoefde. Het instrument is op basis van de bevindingen aangepast en omgezet van een kwantitatief scoringstool naar een online kwalitatief beoordelingsinstrument: de FIAT-Health 2.0. De FIAT-Health 2.0 bestaat uit 13 feitelijke vragen, vragen over de overeenkomst tussen de achterliggende primaire onderzoekspublicatie en de maatschappelijke publicatie en twee open vragen met een beoordeling. De FIAT-Health 2.0 kan beleidsmakers, communicatiemedewerkers en onderzoekers ondersteunen bij het vormen van een systematische en weloverwogen interpretatie van een publiek gerapporteerd cijfer over gezondheid en/of zorg.

Voordat cijfers op een verantwoorde wijze in beleid of praktijk kunnen worden gebruikt, moeten besluitvormers op het terrein van gezondheid en gezondheidszorg deze kritisch onderzoeken en bespreken. Door kritische reflectie op cijfers en hoe dit is ontwikkeld, kan ook worden nagedacht over andere, niet-gemeten aspecten van een onderwerp. Een dergelijke reflectie en discussie zal een beter begrip van statistische methoden en mogelijke toepassingen van cijfers faciliteren. Wanneer besluitvormers moeite doen om cijfers beter te begrijpen, kunnen ze leren deze in de praktijk beter toe te passen. Dit geldt niet alleen voor eenmalige onderzoeksresultaten maar eenzelfde kritische reflectie blijft noodzakelijk voor cijfers die zijn afgeleid van gevalideerde en gestandaardiseerde metingen (i.e. statistieken).

Dit proefschrift biedt een hulpmiddel om kritische reflectie te ondersteunen: de FIAT-Health 2.0. Het instrument is niet bedoeld om een ‘harde beoordeling’ te geven of een cijfer wel of niet kan worden gebruikt, maar helpt de gebruiker om de waarde van een cijfer beter te bepalen en in perspectief te plaatsen. Interpretatie van cijfers is mensenwerk. Elke beleids- of professionele beslissing vereist een afweging waarbij statistische informatie één van de gewichten is. Door de FIAT-Health 2.0 te gebruiken, wordt systematisch gereflecteerd op de kenmerken die relevant zijn voor de interpretatie van cijfers en de bruikbaarheid ervan in verschillende situaties.

Deel II van dit proefschrift (hoofdstuk 5-7) doet verslag van onderzoek naar de mate van verantwoorde rapportage van GZO in wetenschappelijke en maatschappelijke publicaties.

Hoofdstuk 5 beschrijft de omvang en de aard van “twijfelachtige onderzoekspraktijken” “Questionable Research Practices” (QRP’s) in het rapporteren van bevindingen en conclusies in en over wetenschappelijke GZO-publicaties. Samen met 13 deelnemende GZO-instituten in Nederland hebben we QRP’s in de rapportage van bevindingen en conclusies in en over GZO gedefinieerd en een beoordelingsinstrument ontwikkeld voor 35 mogelijke QRP’s. Met behulp van dit beoordelingsinstrument hebben we een willekeurige steekproef van 116 GZO-artikelen beoordeeld. De GZO-artikelen waren geschreven door onderzoekers vanuit de 13 deelnemende GZO-instituten, en zijn in 2016 gepubliceerd in internationale peer-reviewed tijdschriften. We vonden een mediaan van zes QRP's per publicatie. QRP's kwamen het meest voor bij de rapportage van implicaties voor beleid en praktijk, aanbevelingen voor beleid en praktijk, de rapportage van tegengesteld bewijs, rapportage van beperkingen, en conclusies op basis van de resultaten en in de context van de literatuur. We hebben geen verschillen vastgesteld in het totale aantal QRP's per publicatie op basis van verschillende methodologische benaderingen in de onderzochte artikelen, type onderzoek of onderzoeksdesign.

Een verantwoorde wetenschappelijke publicatie is volledig, nauwkeurig en transparant. Ons onderzoek laat zien dat twee onderwerpen met betrekking tot verantwoorde wetenschappelijke rapportage specifieke reflectie vereisen: de vermelding van tegengesteld bewijs en de noodzaak om implicaties en aanbevelingen voor beleid en praktijk op te nemen in een wetenschappelijke publicatie.

Het geconstateerde gebrek aan vermelding van tegengesteld bewijs in wetenschappelijke GZO-publicaties voldoet niet aan de ontwikkelde normen van transparantie in de wetenschappelijke literatuur. Een verantwoorde rapportage in GZO zou altijd een verklaring over tegenstrijdige onderzoeksresultaten dienen te bevatten, zelfs als geen tegengesteld bewijs kan worden gevonden.

Tevens wordt geconcludeerd dat het afzien van het beschrijven van implicaties en aanbevelingen voor het beleid en de praktijk in een wetenschappelijke publicatie de mogelijkheid voor derden creëert om aanbevelingen in het maatschappelijk debat te poneren die niet afdoende gerechtvaardigd worden door de onderzoeksbevindingen. Deze mogelijk verkeerd geïnterpreteerde berichten kunnen onderdeel vormen van het beleidsdebat en zo besluitvormers onjuist beïnvloeden. Het is de verantwoordelijkheid van een gezondheidszorgonderzoeker om te anticiperen op maatschappelijke interesse en gebruik van de bevindingen en hierop ook zelf in een wetenschappelijke publicatie te reflecteren. Daarom dient het opnemen van implicaties en aanbevelingen voor beleid en praktijk een noodzakelijk onderdeel te zijn van een verantwoord gerapporteerde wetenschappelijke GZO-publicatie.

In **hoofdstuk 6** worden inconsistenties tussen wetenschappelijke en maatschappelijke GZO-publicaties beschreven. Daarnaast geeft dit hoofdstuk inzicht in welke QRP's in het rapporteren van berichten en conclusies in wetenschappelijk publicaties gerepliceerd worden in maatschappelijke publicaties. De meeste maatschappelijke publicaties bevatten inconsistenties of gerepliceerde QRP's uit de oorspronkelijke wetenschappelijke publicaties. Gerepliceerde QRPs werden meestal gekenmerkt door twijfelachtig gerapporteerde conclusies, twijfelachtig gerapporteerde beleids- en praktijkaanbevelingen en titels die meer suggereren dan onderbouwd wordt. Er zijn geen substantiële verschillen gevonden in het aantal QRP's tussen maatschappelijke publicaties geschreven door auteurs van de wetenschappelijke publicatie, maatschappelijke publicaties gepubliceerd op websites van onderzoeksinstituten of onderzoeksfinanciers, en maatschappelijke publicaties geschreven door een persoon die niet betrokken was bij de wetenschappelijke publicatie.

Een maatschappelijke publicatie biedt, in tegenstelling tot een wetenschappelijke publicatie, meer vrijheid in de rapportering van onderzoeksbevindingen. De bevindingen die in deel I van dit proefschrift zijn besproken geven aan dat een verantwoorde maatschappelijke publicatie idealiter alle kenmerken bevat die een juiste interpretatie mogelijk maken. Maar omdat maatschappelijke publicaties vaak korte teksten betreffen, passen niet al deze kenmerken in alle maatschappelijke publicaties. Een verwijzing naar de oorspronkelijke wetenschappelijke publicatie lijkt echter een minimumvereiste. Een verantwoorde maatschappelijke publicatie dient niet af te wijken van de bevindingen en discussies in de onderliggende wetenschappelijke publicatie. Maar een kopie van de wetenschappelijke tekst resulteert vaak in een minder begrijpelijk bericht. Beschrijvingen in wetenschappelijke publicaties zijn vaak technisch. Versimpeling van wetenschappelijke bevindingen is veelal noodzakelijk om de aandacht van besluitvormers te trekken.

Bij die versimpeling ontstaat het risico dat onderzoeksresultaten te veel gaan afwijken van de onderbouwing in de wetenschappelijke publicatie. In de beschrijving van de inhoud moet een balans gevonden worden tussen leesbaarheid en consistentie met de wetenschappelijke publicatie. Een verantwoorde maatschappelijke publicatie komt daarom voort uit interactie en discussies tussen onderzoekers en derden, bijvoorbeeld experts op het terrein van wetenschapscommunicatie.

Tot slot, beschrijft **hoofdstuk 7** factoren gerelateerd aan QRP's in wetenschappelijke publicaties. Ervaren druk om maatschappelijke impact te creëren blijkt geassocieerd met een hoger aantal QRP's in de rapportage van bevindingen en conclusies in GZO-publicaties. Specifieke training in het rapporteren van bevindingen en conclusies, evenals bewustwording van belangenconflicten van coauteurs, blijken gerelateerd aan minder QRP's in GZO-publicaties.

Bevindingen uit deze deelstudie kunnen de GZO-gemeenschap helpen bij het ondersteunen van een verantwoorde wetenschappelijke publicatie praktijk. Omdat GZO-instellingen een grote invloed hebben op onderzoekers, hebben zij de mogelijkheid hen te ondersteunen in het schrijven van verantwoorde publicaties. In elke fase van het onderzoeksproces is meer ondersteuning nodig om een verantwoorde wetenschappelijke publicatie te borgen. Trainingsmogelijkheden zijn niet bij alle Nederlandse GZO-instellingen even toegankelijk. Sommige instellingen bieden alleen een basistraining voor wetenschappelijk schrijven, terwijl anderen meer geavanceerde schrijfcursussen aanbieden. Instellingen hebben verschillende structuren voor peer-feedback ingevoerd. Sommige instellingen bieden meerdere mogelijkheden voor peer-feedback in verschillende fases van een onderzoeksproject. Andere instellingen geven onderzoekers de vrijheid feedback te vragen bij collega's wanneer zij behoefte aan steun hebben. Ons onderzoek geeft echter aan dat wanneer peer-feedbackgroepen vrijwillig zijn, onderzoekers deelname soms vermijden. Peer-feedback groepen kunnen alleen succesvol zijn als GZO-instellingen een positieve feedback cultuur ontwikkelen, en onderzoekers mede hierdoor duidelijk voordeel ondervinden bij het schrijven en publiceren van hun manuscripten.

Conclusie

Dit proefschrift biedt inzicht in de kenmerken die nodig zijn om publiek gerapporteerde cijfers over gezondheid en gezondheidszorg te interpreteren, en biedt een methode voor individuele onderzoekers, wetenschapscommunicatie experts/journalisten en besluitvormers om met deze kenmerken tot een weloverwogen oordeel te komen over een cijfer door gebruik van de FIAT-Health 2.0. Als individuele onderzoekers en wetenschapscommunicatie experts/journalisten het instrument gebruiken om de publieke rapportage van cijfers te verbeteren, en wanneer besluitvormers het instrument gebruiken om cijfers beter te interpreteren, kan het verantwoorde gebruik van cijfers worden bevorderd. Twijfelachtige onderzoekspraktijken (QRP's) in het rapporteren van bevindingen en conclusies komen vaak voor in peer-reviewed internationale wetenschappelijke GZO- publicaties van Nederlandse onderzoeksinstituten. Met name het rapporteren van tegengesteld bewijs en aanbevelingen en implicaties voor beleid en praktijk, kunnen worden verbeterd. Naast het voorkomen van inconsistenties tussen wetenschappelijke en maatschappelijke publicaties, worden QRP's in het rapporteren van bevindingen en conclusies van wetenschappelijke publicaties vaak overgenomen in maatschappelijke publicaties. Om de huidige staat van de wetenschappelijke rapportage in GZO te bevorderen, zou de GZO-gemeenschap ondersteunende maatregelen moeten nemen in de vorm van schrijftraining gericht op de rapportering van bevindingen en conclusies. Daarnaast dienen onderzoekers alert te zijn op mogelijke belangverstrengeling van coauteurs. Tot slot zou de leiding van GZO-instituten, onderzoeksfinanciers en editors van wetenschappelijke tijdschriften beleid, om druk op onderzoekers te verminderen op het creëren van maatschappelijke aandacht ,verder dienen door te voeren.

Het vraagteken uit de titel van dit proefschrift “Open to Interpretation?” zal blijven. Echter, de hier gepresenteerde inzichten kunnen bijdragen aan een verantwoorde (GZO-) onderzoekspraktijk en correcte interpretatie van cijfers en onderzoeksbevindingen.

DANKWOORD

Na vier jaar vol met interviews, focusgroepen, vragenlijsten, checklists, beleidsdocumenten, beoordelingsinstrumenten, analyses en schrijfwerk is het zover: mijn proefschrift staat! Al dat werk had ik nooit kunnen doen zonder de steun van veel vrienden, familie en collega's. Aan hen zou ik aan het einde van dit avontuur nog wat extra aandacht willen besteden.

Allereerst wil ik graag mijn **promotor, prof. dr. Niek Klazinga** bedanken. Niek, jouw inzicht in gezondheidszorgonderzoek is onvoorstelbaar. Jij hebt mij in het gezondheidszorgonderzoek geïntroduceerd en laten zien hoe de wereld van de wetenschap werkt. Keer op keer wist je feilloos de kern van mijn werk naar boven te halen, de juiste vragen te stellen en het belang van ons onderzoek te schetsen. De promotieoverleggen met Dionne en Michael bij jou thuis en de omgeving die jij creëerde waren een enorme ondersteuning. Ik kan mij geen betere promotor voorstellen.

Daarnaast wil ik graag mijn **copromotoren dr. Dionne Kringos en dr. Michael van den Berg** bedanken. Dionne, dankzij jou heb ik al deze jaren naar dit doel toe kunnen werken. Jij hebt mij geleerd om focus te ontwikkelen, een onderzoekshouding aan te nemen, taken te balanceren en mijn eigen werk op waarde te beoordelen. De afgelopen jaren heb ik heel goed met jou samengewerkt. Ik kon er altijd op rekenen dat je structuur kon vinden in mijn ideeën en dat we die ideeën samen konden aanscherpen. Dit proefschrift had hier niet gelegen zonder jouw begeleiding.

Michael, je hebt mij laten zien hoe cijfers en letters op de juiste plek komen te staan. Van jou heb ik geleerd om zorgvuldig te werken, vast te houden aan wetenschappelijke methoden en niet bang te zijn om moeilijke keuzes te maken in mijn werk. Ik wil je enorm bedanken voor alle spontane discussies op de gang bij het RIVM. Ook nadat je naar Parijs vertrok, kon ik altijd op jou rekenen via Skype.

Ik wil bovendien graag de leden van mijn **promotiecommissie** hartelijk bedanken, prof. dr. A.E. Kunst, prof. dr. L.M. Bouter, prof. dr. A.J. Pols, prof. dr. A. Verhoeff en prof. dr. D. Ruwaard, voor het lezen en beoordelen van mijn proefschrift.

Els, mijn eeuwige kamergenootje en paranimf, dankjewel voor alle thee, gesprekken, kantoor oefeningen en gezelligheid de afgelopen jaren op kantoor, in Stockholm, in cafés en thuis op de bank. Jij hebt mij door al deze jaren heen geloodst en ik ben je enorm dankbaar dat ik deze tijd samen met jou af kan ronden. **Marjonneke,** jij hebt die laatste maanden schrijven zoveel makkelijker en gezelliger gemaakt. **Tessa and Joko,** my co-authors, doing research was the most fun when I was doing it together with you. **Wahyu, Ben, Xanthe, Anke, and Laxsini,** thank you for being such wonderful and fun ~~distractions~~ roommates. Thanks to you I could laugh every day. I already miss those glühwein parties as tradition!

Paulien, Michael, Mirthe, Gertrude, and Sigrid, I'd like to thank you especially for sharing these 4 years with me. EPH conferences, APH seminars, CaRe-days and simply every day at the department were made special because of you.

Ik wil ook graag **Syeldy**, **Danae, Wilco, Noukie, Lisa, Renee, Adriëtte, Karlijn, Mirte, Frederique**, **Marieke, Marianna en Marloes** bedanken voor alle lunches, koffiemomentjes, borrels en wandelingen.

Janine, Myrthe, Aydin en Anke, het was ontzettend leuk om de QoC PhD-werkgroep met jullie op te zetten! **Noor, Henriëtte en Nita**, dank jullie wel voor alle secretariële ondersteuning de afgelopen jaren! En **Joost en Rowan,** het was ontzettend leuk om jullie te mogen begeleiden! En **Lize,** dankjewel voor de prachtige binnenkant van dit proefschrift.

Ik wil ook graag mijn oud-collega's bij het **Centrum Gezondheid en Maatschappij van het RIVM** bedanken. De eerste jaren van mijn PhD hebben jullie mij laten zien hoe cijfers verantwoord gecommuniceerd worden naar een groter publiek. Ik heb het altijd ontzettend leuk gevonden om iedere week naar Bilthoven te komen en met jullie samen te werken!

Natuurlijk kan ik mijn nieuwe collega's bij **Equalis** niet vergeten! Jullie hebben mij zo ontzettend goed ontvangen dat ik mij al welkom voelde voordat ik begon met werken. In deze eerste maanden heb ik al waanzinnig veel van jullie geleerd, gelachen en ondersteuning ontvangen. Ik ben enorm dankbaar dat ik mijn carrière bij jullie kan voortzetten.

Mijn proefschrift zou er niet gekomen zijn zonder de steun van mijn familie. **Annette**, je hebt mij al zo vaak verhuisd. Ik ben enorm blij dat je mijn paranimf wilt zijn en zo mij ook met deze overgang helpt. Zonder jouw hulp ben ik nergens. **Papa en mama**, dankjewel dat ik altijd bij jullie terecht kan en dat ik mijn eigen kantoortje kon inrichten in de woonkamer. Dankzij jullie kon ik dit proefschrift in alle rust afschrijven.

Dear **Alyssa**, thank you so much for designing the cover of this book. You magically found a way to make the cover fit the topic of my thesis in a beautiful minimalistic style.

My Patrick, when we met at the beginning of both our PhD's, we had a short but somehow interesting conversation about statistics. We didn't know it then, but we would have many more. You showed me an enthusiasm for science, sharpened my arguments (frustratingly) and talked me through difficult moments. Even from the other side of the world you provided the balance that I needed. Your love motivated me to keep going and be proud of my work. I'm so happy that I can share this experience with you.

ABOUT THE AUTHOR

Reinie Gerrits was born on February 11th, 1992 and raised in Meerlo, the Netherlands. In 2010, Reinie completed high-school at Dendron College in Horst. She continued her education at Maastricht University where she obtained a bachelor of science degree in European Public Health. Further, she completed a minor in International Relations at Malmö University in Sweden and wrote her bachelor's thesis at the Emergency Medical Services in the Capitol Region of Denmark in Copenhagen, Denmark.

In 2013, Reinie moved to Stockholm, Sweden and obtained her master's degree in Public Health Sciences with a specialization in Health Economics, Policy and Management at the Karolinska Institute. During this time, Reinie was also part of the Public Health Sciences Board in the Medicinska Föreningen student organization, student representative in the Programme Council of her masters' specialization, and one of the inaugural members of the Swedish Network for International Health (SNIH). Additionally, Reinie completed a "Doing Business in China" summer course at Fudan University in Shanghai, China. She graduated from the Karolinska Institute in 2015 after completing her master's thesis, "Patient online access to medical records in the Netherlands, an explorative study on the interplay of policy maker and care provider initiative".

After leaving the Karolinska Institute, Reinie worked toward a Ph.D. in the Public Health department at Amsterdam UMC under the supervision of Prof. Dr. Niek Klazinga, as well as at the Netherlands Institute of Public Health and the Environment (RIVM). She was involved in several projects, including "Statistics on Health and Healthcare" at the Institute in Bilthoven in 2017 and "Fostering Responsible Research Practices", a project funded by the Netherlands Organization for Health Research and Development (ZonMw).

Outside of her research, Reinie coordinated and led courses in Medical Informatics and Evidence Based Practice along with short courses at the Netherlands Institute for Health Sciences. As a council member at the Amsterdam Public health – Quality of Care Research program, a member of the social activity committee at the department of public health, and a founder of the PhD-junior council at the Quality of Care research program, Reinie has dedicated herself throughout her career to creating a supportive social environment for her colleagues.

Reinie's primary career objective is to improve healthcare practice through the responsible conduct and implementation of research. Since December, 2019, she has continued working towards this goal as a health consultant at Equalis Strategy and Modeling.

PORTFOLIO

Name PhD student:	Reinie G. Gerrits
PhD period:	November 2015 – February 2020
Names PhD supervisors:	Prof. Dr. N.S. Klazinga Dr. D.S. Kringos Dr. M.J. van den Berg

	Organisation	Year	ECTS[a]
General courses			**0.3**
Web of Science	AMC	2016	0.1
Endnote	AMC	2016	0.1
Presentation Skills	AMC	2016	0.1
Specific courses			**1**
International Comparison of Health Care Systems	NIHES	2016	1
Seminars, workshops and master classes			**3.2**
Seminar: "Anatomische les" (2x)	AMC	2016-2017	0.2
Seminar: Quality indicators for hospital care	Erasmus MC	2016	0.1
Workshop: House of Cards	AMC	2016	0.1
Colloquium "Centrum Kennisintegratie Volksgezondheid en Zorg" (x6)	RIVM	2016 - 2017	0.6
Amsterdam Public Health research institute: Quality of Care meetings (x3)	APH	2016	0.6
Netherlands School of Primary Care Research CaRe Days in Maastricht	CaRe	2017	0.5
Netherlands School of Primary Care Research CaRe Days in 's-Hertogenbosch	CaRe	2018	0.5
The Netherlands Organisation for Health Research and Development - Program meeting "Fostering Responsible Research Practices" on Research Waste	ZonMw	2018	0.2
WHO seminar on primary healthcare system Malta	AMC	2015	0.4
Presentations			**2.2**
Colloquium "Centrum Kennisintegratie Volksgezondheid en Zorg" (x2)	RIVM	2016 - 2017	0.5
Public Health department presentations (3x)	AMC	2015- 2019	0.1
Peer-review group meetings (x2)	AMC	2018	0.1
Workshop employee-day Ministry of Health, Welfare, and Sports	VWS	2017	0.5
Workshop junior council Quality of Care: Argumentation	Amsterdam Public Health research institute	2018	0.5
The State of Health and Healthcare: Scientific Advisory Board presentation	RIVM	2017	0.5

	Organisation	Year	ECTS[a]
(Inter)national conferences			**12**
EUPHA Vienna + presentation (x2)	EUPHA	2018	2
EUPHA Stockholm + presentation	EUPHA	2017	1.5
FUSE Newcastle	FUSE	2016	0.75
“Bessensap” Amsterdam (x2)	NWO	2016 – 2017	0.5
Amsterdam Public Health 1st annual meeting + poster	APH	2016	0.75
Professional conference science communication The Hague + workshop	WTC	2017	2
Professional conference science communication Amsterdam + workshop	WTC	2019	2
NRIN conference + presentation	NRIN	2018	1
NCVGZ Rotterdam	NCVGZ	2016	0.5
NCVGZ Amersfoort + presentation	NCVGZ	2017	1
Teaching			**11.35**
Master thesis intern, MSc Management, Policy Analysis & Entrepreneurship in Health sciences	VU	2017	1.6
Literature thesis intern, MSc Management, Policy Analysis & Entrepreneurship in Health sciences	VU	2017	0.5
Lecture: Short course Health Services Research and Practice	NIHES	2017	0.25
Lecture: BSc Medical Information Science (x2)	UVA	2017 - 2018	0.5
Practical: BSc Medicine workgroup Quality of Care (x 7)	UVA	2016 - 2017-2018 -	0.5
Coordination of BSc Medical Information Science course Volksgezondheid en gezondheidszorg	UVA	2017	5
Coordination of MSc Evidence Based Practice course Health Care Policy Evaluation	UVA	2016	2
Coordination Short course Health Services Research and Practice	NIHES	2016 - 2017	0.5
Coordination Short course ‘International Comparison of Health Care Systems’	NIHES	2016 - 2017	0.5
Other activities			2.5
Amsterdam Public Health Quality of Care: Program council	APH-QoC	2017 - 2018	1
Amsterdam Public Health Quality of Care: Junior council	APH-QoC	2017-2018	1
Department of Public Health social committee	AMC	2017-2018	0.5
Grants			0.3
ZonMw research grant co-author “Fostering Responsible Conclusions and Messages in Health Services Research”	ZonMw	2017	0.3
Total ECTS			**32.85**

[a] *ECTS = European Credit Transfer and Accumulation System; 1 ECTS = 28 hours workload*

LIST OF PUBLICATIONS

Publications included in this thesis

Gerrits RG, Kringos DS, van den Berg MJ, Klazinga NS. Improving interpretation of publicly reported statistics on health and healthcare: the Figure Interpretation Assessment Tool (FIAT-Health). Gerrits RG, Kringos DS, van den Berg MJ, Klazinga NS. Health Res Policy Syst. 2018 Mar 7;16(1):20. doi: 10.1186/s12961-018-0279-z.

Gerrits RG, Jansen T, Mulyanto J, van den Berg MJ, Klazinga NS, Kringos DS. Occurrence and nature of questionable research practices in the reporting of messages and conclusions in international scientific Health Services Research publications: a structured assessment of publications authored by researchers in the Netherlands. BMJ Open. 2019 May 15;9(5):e027903. doi: 10.1136/bmjopen-2018-027903.

Gerrits RG, van den Berg MJ, Klazinga NS, Kringos DS. Statistics in Dutch policy debates on health and healthcare. Health Res Policy Syst. 2019 Jun 3;17(1):55. doi: 10.1186/s12961-019-0461-y.

Gerrits RG, Klazinga NS, van den Berg MJ, Kringos DS. Figure Interpretation Assessment Tool-Health (FIAT-health) 2.0: from a scoring instrument to a critical appraisal tool. BMC Med Res Methodol. 2019 Jul 23;19(1):160. doi: 10.1186/s12874-019-0797-6.

Gerrits RG, van den Berg MJ, Kunst AK, Klazinga NS, Kringos DS. Fostering responsible research practices in societal publications of Health Services Research. Submitted.

Gerrits RG, Mulyanto J, Wammes JD, van den Berg MJ, Klazinga NS, Kringos DS. Individual, institutional, and scientific environment factors associated with questionable research practices in the reporting of messages and conclusions in scientific Health Services Research publications Submitted.

Publications not included in this thesis

Essén A, **Gerrits R**, Kuhlmann E. Patient accessible electronic health records: Connecting policy and provider action in the Netherlands. Health Policy and Technology 2017; 6(2): 134-41.

Essén A, Scandurra I, **Gerrits R**, et al. Patient access to electronic health records: Differences across ten countries. Health Policy and Technology.

Gerrits RG, Kringos DS, van den Berg MJ, Klazinga NS. [Rapport] De totstandkoming van de Figure Interpretation Assessment Tool-Health (FIAT-Health). Amsterdam: AMC, Afdeling Sociale Geneeskunde; 2016.

Gerrits RG, Kringos DS, van den Berg MJ, Klazinga NS. [Rapport] De betrouwbaarheid, validatie, en gebruiksvriendelijkheid van de FIAT-Health. Amsterdam: AMC, Afdeling Sociale Geneeskunde; 2017.

Gerrits RG, Kringos DS, van den Berg MJ, Klazinga NS. Improving numerical literacy for policy makers: the Figure Interpretation Assessment Tool (FIAT): Reinie Gerrits, European Journal of Public Health, Volume 26, Issue suppl_1, 1 September 2016, ckw171.069, https://doi.org/10.1093/eurpub/ckw171.069

Kringos DS, van den Berg MJ, Klazinga**, Gerrits RG**, The use of figures in the recent Dutch policy debate on health and healthcare: a discourse analysis: Reinie Gerrits, European Journal of Public Health, Volume 27, Issue suppl_3, November 2017, ckx187.341, https://doi.org/10.1093/eurpub/ckx187.341

Gerrits RG. Box 3.3 The Figure Interpretation Assessment Tool – Health 1.0. In: Structuring Health Information: Frameworks, Models and Indicators: Climbing the Information Pyramid. January 2019 In book: Population Health Monitoring. DOI: 10.1007/978-3-319-76562-4_3.