NB: Part 1 includes an extended abstract to the conference presentation. Part 2 includes a working paper which features much of the background for the abstract and conference presentation. This is entitled ‘Positivism, Postpositivism and Evaluation in Health: The Case for a Qualitative Approach to the Evaluation of Policy and Governance’.
New Public Management, Evidence and Tensions between Centralisation and Decentralisation: the Case of Diabetes Services

Extended abstract

The rise of the New Right during the late 1970s and 1980s saw increasing acceptance of the superiority of the state over markets in the coordination of economic activity. Part of the argument for the superiority of markets hinged upon established accounts of the incentives that markets facilitate, which purportedly ensure individuals work in the interests of others, even though they may have interests of their own. Yet part of the argument hinged upon Frederick Hayek’s theorisation of the economic problem as one of facilitating the successful utilisation of knowledge dispersed throughout society. Hayek provided an influential account of the market system as a decentralised mechanism which facilitates coordination via price signals, which not only provide signals of consumer demand but enable producers to utilise their own knowledge and expertise in the creation of consumer items. That account would inform New Right criticisms of state involvement in the economy and the welfare state, said to be inefficient, rigid and mechanistic (Wilson, 2004: 51).

The cogency of Hayek’s criticism of centralised planning is attested by its acceptance across the political spectrum. The collapse of the Soviet Union is widely held to be explained by Hayek’s theory: central planners had seemingly failed to acquire the knowledge of consumer preferences and technologies required to efficiently plan and deliver goods and services in a complex modern economy. However, despite this apparent agreement that central planning is untenable, a central paradox of market reforms to public services today is that in many respects public service governance has never been more hierarchical. Commentators in public administration note a similarity between public service governance today and the system of central planning which existed in the Soviet Union (Bevan and Hood, 2006; Hood, 2006; Ann et al., 2009; Propper et al., 2008).

Indeed, via the application of New Public Management (NPM), governments have sought to mimic markets in public services. In some respects, NPM decentralises decision-making: large public bureaucracies are disaggregated into smaller units which are to specialise on core services, purchaser-providers splits are introduced and organisational units compete for contracts to deliver public services. In theory, the combination of disaggregation and competition should facilitate a consumer orientation and enhance efficiency. However, in other respects, NPM centralises decision-making. Performance management protocols typically specify strict standards of service delivery, performance is monitored in relation to targets and market contracts are specified by the centre. Additionally, decision-making is increasingly made on the basis of evidence. Indeed, Evidence-Base Policy is a further key feature of NPM and seeks to ensure that decisions at all levels of policy and governance are evidence-based. This would purportedly improve performance by lessening reliance on ‘fallible’ professional judgement and ‘ideological’ political judgement (Banks, 2009; Chalmers, 2003; Thornton, 2006).

The combination of these approaches have proved controversial, ranging from concern they are a precursor for outright privatisation to concern that public sector organisations today are ill-equipped to solve complex policy problems (Head and Alford, 2013). Alluding to Hayekian criticisms of central planning, some argue that the use of centralised performance management protocols and standardised contracts inhibits the scope for local actors to
utilise local knowledge and collaborate in the pursuit of policy objectives (Head and Alford, 2013: 9).

This presentation explores this tension between decentralisation and centralisation through a focus on health and in particular diabetes services. It draws upon ‘postpositivist’ policy analysis to critique dominant methods of evaluation in health and presents findings of some empirical research into diabetes services which suggest current policy and governance arrangements are not striking the right balance between centralisation and decentralisation: there is significant contestation over key targets in diabetes care and concern that marketisation is fragmenting care pathways. Additionally, recent reforms to commissioning appear to frustrate the effective utilisation of existing medical knowledge and expertise.
Abstract

In recent years, the 'postpositivist' turn in the policy sciences has informed some significant criticisms of quantitative approaches to evaluation on the basis that they fail to address issues of complexity and contestation. However, positivism remains the dominant research framework in evaluation in health, evident in the reliance on a range of quantitative and experimental methodologies.

Drawing upon postpositivist perspectives in the philosophy of science, this paper criticises the influence of positivism over the field and makes a case for a qualitative approach to evaluation. It is argued that positivism is not only philosophically problematic but is leading to potentially significant distortions in decision-making. At the level of policy, where decision-makers face choices between particular interventions, services and pathways of care, quantitative evaluations appear to underestimate the potential of holistic forms of care and overstate the potential of pharmacological therapies, leading to their oversupply. Similarly, some examples of quantitative evaluation overstate the performance potential of New Public Management reforms, including performance management and marketisation. Evaluations at this level tend to take the form of retrospective observational studies in which statistical techniques are used track performance over time. However, these do not provide the level of detail required to capture the full impacts of policy. The paper ends with proposals for a postpositivist, qualitative approach to evaluation.

1. Introduction

Since the emergence of policy analysis in the 1960s, there has been extensive debate about what constitutes acceptable forms of evidence. This debate aligns with the broader ‘paradigm wars’ in the social sciences (Dryzek, 1994; Fischer, 2003; Stone, 2001). Positivism has had significant influence over policy analysis and can be identified in recent calls for the greater use of evidence in decision-making. The Evidence-Based Policy movement seeks to improve the performance of public services by ensuring that all services have been rigorously evaluated. It is anticipated that evidence-based decision-making can replace “ideological” political processes (Banks, 2009) and “fallible” professional judgment (Thornton, 2006: 2; Chalmers, 2003). The approach can be linked to the use of managerialism and marketisation in New Public Management (NPM) because evaluation provides a way to identify best practices which can subsequently inform the content of performance management protocols and market contracts.

However, the scientific and economic methodologies promoted in Evidence-Based Policy have been criticised, particularly in health where there is frequent contestation over evidence. More broadly, NPM has itself been criticised for reinforcing a ‘silod’ approach to public service delivery which fails to address complex policy problems (Head and Alford, 2013: 9). Criticisms of Evidence-Based Policy and NPM today resemble wider criticisms of positivism in policy analysis. Frank Fischer criticises the influence of positivism over policy analysis. According to Fischer, positivism has led policy analysis to take on an increasingly technocratic orientation which suppresses the ethical and political dimensions of decision-making (Fischer, 2003). A related point is that positivist policy analysis downplays the uncertainties and the complexity of means-ends interrelationships involved in these
There is a need to engage critically with existing methodological approaches to evaluation and develop alternatives that are more attendant to issues of complexity and contestation.

The present paper critiques positivist understandings of evaluation and evidence in health, while also making a case for a postpositivist alternative. Section 2 introduces the positivist research paradigm and some criticisms of it. The section explores a number of alternative research frameworks (postmodernism, interpretivism, critical realism), each of which make significant contributions to debates about method. A case is made for methodological pluralism in research and evaluation.

Section 3 introduces evaluation in health, identifying the main research areas and evaluative methodologies that are applied in the research. A heuristic which features in the section is a distinction between policy and governance. This serves to elucidate the nature of the challenges confronted by decision-makers at different stages of the policy process: at the level of policy, decision-makers face choices between interventions and services which seek to solve particular policy problems; at the level of governance, decision-makers face choices between organisational forms and governance approaches which affect the delivery of policy at a local level. Corresponding to this distinction, specific forms of evaluation have been developed to inform decision-making in health: health technology research and health services research. Both research areas are significantly influenced by positivism, placing emphasis on the importance of quantitative and experimental methodologies.

Section 4 draws out some of the implications of the dominance of positivism for health policy and governance: at the level of policy, dominant forms of evaluation appear to understate the potential of holistic interventions and overstate the potential of pharmacological interventions, leading to their oversupply. At the level of governance, some examples of quantitative evaluation overstate the performance potential of New Public Management reforms, including performance management and marketisation. The paper ends with proposals for a postpositivist approach to evaluation involving empirical research.

2. Positivism and its Critics

There has been significant methodological debate in the social sciences since its inception, in the 18th century. This is often characterised as a debate or "war" between those who believe the social sciences are, or ought to be, the same as the natural sciences and those who believe the two are fundamentally different. This simple demarcation no longer holds as there is increasing recognition of the need for methodological pluralism in social research and indeed evaluation. Nevertheless, positivism has had a significant if often unacknowledged influence over all forms of social research.

What exactly is implied when researchers adopt, implicitly or explicitly, a positivist framework? A key assumption of positivism is that objective knowledge is possible, provided rigorous methods are applied. The purpose of research is to establish facts about the world that are independent of the researcher and their values, at least in methodologically rigorous research. Values in this way are potentially separable from facts. Facts, for their part, take the form of either causes or effects and can be ascertained through observation and measurement. Once data has been collected, researchers can examine relationships between facts through statistical analysis, on the assumption that reality is characterised by

---

1 Defining positivism is somewhat precarious because few researchers consciously describe themselves as ‘positivists’, with the term read onto some peoples’ work as a “term of abuse” (Ishiyama and Breuning, 2010: 461).
cause/effect relations\(^2\) (Hagan, 2012: 177). This process enables the testing of hypotheses which if verified assume the status of scientific laws\(^3\).

While the following section highlights the influence of positivism over research and evaluation in health, it is interesting to first note the significant criticisms that have been made of the framework in the philosophy of science and social sciences. Table 1 presents positivism and three distinct perspectives that are critical of the framework, aligned in order of the extent they uphold the possibility of objective and generalisable knowledge.

\(^2\) This is the so-called ‘successionist’ conception of causation based on the following form: if B follows A, then A causes B (McLaughlin and Newburn, 2010: 205; see also Pawson, 2008).

\(^3\) There are differences within positivism regarding the role of theory in the development of hypotheses and the status of hypotheses once they are tested. Karl Popper, for example, allows an extensive role for prior theory in the development of hypotheses and has argued that rather than proceed on the basis of verification – implied by the positivist notion of ‘laws’ – science proceeds on the basis of falsification: all that science can establish is “the creation of, as yet, unfalsified laws” (Gray, 2009: 22).
## Research Frameworks

<table>
<thead>
<tr>
<th>Methodology</th>
<th>Positivism</th>
<th>Critical Realism</th>
<th>Interpretivism</th>
<th>Postmodernism</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Quantitative: experiments, surveys and statistical analysis</td>
<td>Mixed methods: the correct research depends on the subject matter</td>
<td>Qualitative: in-depth interviews, participant observation, action research and discourse analysis</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Epistemology</th>
<th>Positivism</th>
<th>Critical Realism</th>
<th>Interpretivism</th>
<th>Postmodernism</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Objective knowledge is possible provided that rigorous methods are applied</td>
<td>Knowledge is always relative but scientific methods provide some grounds to evaluate truth claims</td>
<td>Knowledge is relative but intersubjective agreement provides grounds to evaluate truth claims</td>
<td>Knowledge is relative and power determines which truth claims prevail</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ontology</th>
<th>Positivism</th>
<th>Critical Realism</th>
<th>Interpretivism</th>
<th>Postmodernism</th>
</tr>
</thead>
<tbody>
<tr>
<td>A single reality exists, characterised by cause/effect relations</td>
<td>A single reality exists but is complex and emergent: cause/effect relations identified in one context may not generalise to another</td>
<td>Reality is at least partly socially constructed and multiple realities exist</td>
<td>Reality is entirely socially constructed and multiple realities exist</td>
<td></td>
</tr>
</tbody>
</table>
Postmodernism is the framework that is most hostile to positivism, refuting any notion of objective knowledge and indeed causal relations which underpin scientific theories (Atabor, 2014). Postmodernism is influenced by historical accounts of the sciences which emphasise their embeddedness in the cultures in which they are practiced and the role of established scientific theories and frameworks in determining what is accepted as true at any point in time (Kuhn, 2012). Developments in the natural sciences themselves have also proved influential. Quantum theory and chaos theory suggest that causes can have multiple effects and vice versa, while small changes in any natural or social system can have large and often unexpected effects. This, in turn, undermines the positivist view of an orderly natural and social world which is amenable to manipulation and control.

This questioning of positivist epistemology and ontology has informed the claim of postmodernists that values are intimately connected to facts and the pursuit in research of a singular, generalisable truth is futile. Furthermore, postmodernists claim that truths are determined more by power than reasoned inquiry. There are no grounds to evaluate truth claims because any criteria are always partial and favour one individual or group over another. However, while the framework provides a valuable check on the positivist assumption that scientific methods provide a neutral mechanism to adjudicate between truth claims, the framework poses the question of whether it is legitimate to even attempt to identify solutions to policy problems through research. If values are intimately connected to facts and truth is determined by power, the entire edifice of scientific inquiry collapses and it is impossible to have rational policymaking.

The interpretivist critique of positivism emphasises differences between society and nature. Society, unlike nature, is not characterised by cause/effect relations but is constituted by human beings who ascribe meanings to their actions and to social events. The fundamentally different makeup of society implies that different methods are required to study it for the natural and social sciences pertain to different domains. Researchers are also part of the social world and cannot simply observe it as a phenomenon external to them. For these reasons, interpretivists emphasise the importance of qualitative methods in social research. To establish truths about the social world requires that researchers drop prior assumptions and categories and establish a dialogical relationship with their subject matter, exploring how social phenomena appear in the consciousness of research subjects (Englander, 2012). In policy analysis, these arguments have been influential in the development of qualitative approaches to evaluation which proceed by identifying and analysing interpretations or ‘framings’ of policy problems, which feature potentially conflicting values and different understandings of how they might be solved (Rein and Schon, 1996; Yanow, 1997).

Still, interpretivism has been criticised for making too much of the distinction between society and nature. While society does have subjective and intersubjective elements to it which may

---

4 Thomas Cook argues that reality is not made up of simple cause/effect relations, but causes can have multiple effects (and vice versa) and intervening variables complicate efforts to test hypotheses. This is particularly true of society:

Human relationships are more like pretzels than single-headed arrows from A to B ... more like convoluted multivariate statistical interactions than simple main effects (Cook, 1985: 25).

5 Indeed, postmodernism has been sharply criticised in recent years for this very reason. The framework leaves a weak basis for progressive research agendas (Collier, 1994) and is also self-refuting since it depends upon truth claims to deny the possibility of truth (Putnam, 1982).
be best explored through qualitative research, human actions still have consequences which
take the form of cause-effect relations. Furthermore, society is constituted by relatively
durable social structures. Quantitative research can be useful in identifying the presence of
social structures which may not be immediately apparent to the social actors who reproduce
them, such as class or unequal gender relations. Rather than argue for a fundamentally
different kind of science for the social sciences, therefore, a balance of quantitative and
qualitative research may be more appropriate (Alvesson and Skoldberg, 2009: 4).

An influential framework that has emerged in recent years is critical realism, which seeks to
reconstruct the social and natural sciences for progressive purposes. Critical realism
incorporates elements of the postmodernist and interpretivist critique of positivism but also
upholding the possibility of progress and emancipation through the acquisition of objective
knowledge. For critical realists, recognising that actors have different interpretations of reality
does not imply that reality itself is socially constructed, for that is to confuse epistemology
with ontology. Furthermore, exploring actors’ different interpretations of reality through
qualitative research improves our understanding of objective reality and can be
complementary to quantitative research (Forsyth, 1998).

A further dimension of critical realism is its critique of positivist understandings of the natural
sciences. In his influential book ‘A Realist Theory of Science’, Roy Bhaskar criticised the
‘successionist’ conception of causality but nevertheless upholds the possibility of scientific
explanations of cause-effect relations. Bhaskar’s criticism of positivism hinges on a
distinction between the ‘closed systems’ of scientific experiments and the ‘open systems’ of
the natural and social realms. Science proceeds by engineering ‘closed systems’ in order
that variables can be manipulated and controlled, allowing for the identification of causal
relationships. While this approach facilitates the generation of scientific knowledge, research
findings will to a certain extent be specific to the conditions of the experiment. For this
reason, Bhaskar argues that it is never possible for scientists to uncover laws that are
generalisable to all contexts; instead, scientists can only seek to uncover tendencies which
may or may not occur in open systems (Bhaskar, 2008).

This critical realist paradigm provides a flexible epistemological perspective and highlights
the potential for methodological pluralism in research and evaluation. It rejects attempts to
prioritise some method over another, for the appropriate method will be determined by the
research topic and research question. This flexible approach is notably absent in health,
where quantitative and experimental methodologies are prioritised irrespective of the topic of
research.

3. Evaluation and Evidence in Health

Evidence-Based Medicine and Policy emerged during the 1990s and was effectively
institutionalised following the creation of the National Institute for Clinical Excellence (NICE)
by New Labour in 1999, after which evaluation and evidence would become increasingly
central to policy-making. The approach seeks to improve the performance of health services
by ensuring that the most efficient and effective forms of care are delivered. A key aspect of
this is setting out clear standards of evidence which encourage the application of certain
methods over others and can be used in the appraisal of existing evidence (Habour and
Miller, 2001; Ho et al., 2008).

However, there has been significant debate over the significance which the field assigns to
quantitative methodologies. Just recently, a group of 76 academics contacted the British
Medical Journal to complain about the journal’s increasingly apparent failure to publish
qualitative research. In a reply, a spokesperson stated that “qualitative studies are an
extremely low priority for the BMJ" and the journal has to accept and reject studies on the basis of their rigour, practical value and interest to general clinical readers (Greenhalgh et al., 2016: 3). This assumption that quantitative research is more rigorous and useful than qualitative research extends into clinical research and health policy research.

Evidence-Based Medicine and Policy can be divided into health technology research and health services research, which correspond to the distinction between policy and governance discussed above. Health technology research involves evaluations of interventions and services such as diagnostic techniques, surgical procedures and pharmacological therapies. In this area of research, the Randomised-Control Trial (RCT) is considered the “gold standard” of research and is situated at the top of an evidence “pyramid” (Ho et al., 2008) or “hierarchy” (Habour and Miller, 2001). RCTs proceed by randomly allocating research participants to either an experimental or a control group. Participants in the experimental group are exposed to whatever intervention is under evaluation and outcomes in both groups are measured. Once data has been collected, statistical tests can be applied to establish the degree of certainty that the outcomes identified in the experimental group can be attributed to the intervention (Campbell and Stanley, 1966). Below this in the evidence “hierarchy”, observational studies evaluate services and interventions in practical environments where randomisation and controlled experimentation is not possible.

While RCTs are considered the “gold standard” of primary research, even superior to this is the systematic review which synthesise data from multiple research studies. Systematic reviews, which ideally synthesise research findings of RCTs, provide a more comprehensive evaluation than could ever be the case with a single research study. The Cochrane Collaboration provides highly influential systematic reviews on health interventions which are used in medical and policy decisions (Sackett and Wennberg, 1997).

Together, the RCT and the systematic review are considered the surest way to establish the efficacy of an intervention. However, a major criticism of these approaches is their tendency to focus only on medical outcomes, either in terms of hard endpoints, such as mortality and survival rates, or disease-specific effects that are particular to certain conditions and diseases6. There is a danger that other non-medical criteria come to be suppressed, such as the quality of life effects of treatments or their cost.

To overcome this problem, a form of cost/benefit analysis has been developed in healthcare to provide a more comprehensive evaluation of treatment options, known as Cost-Utility Analysis (CUA). CUA combines clinical data with data on the subjective preferences of patients. Patients with a condition are asked to indicate where they perceive themselves to be on a utility scale between a point below zero, zero and one, where below zero represents a situation that is worse than death, zero represents death and one represents perfect health. Through this procedure, all conditions, whether mental or physical, can be assigned a utility rating, which in turn can be used to evaluate the quality of life effects of treatment. If clinical data suggests that a treatment contributes ten years of extra life but the quality of those extra years scores a 0.5 utility rating, then the treatment contributes 5 QALYs. Furthermore, the QALY rating of a treatment can be divided by its cost to establish the price of each QALY, providing a further point of comparison. In theory, the evaluation of treatments in this way enable decisions to take into account a variety of criteria, including efficiency, the quality of care (from a medical perspective) and the quality of life (from the patient perspective) (Jones, 2012; Whitehead and Ali, 2010).

---

6 Examples of disease specific effects include lung capacity for lung disease.
While the next section explores the major criticisms of established health technology research, one limitation that is recognised in the literature is that health technology evaluations only seek to inform policy decisions and refrain from addressing evaluative questions pertaining to the wider organisational and governance environment. This is potentially problematic because the treatments and services that work under experimental conditions – or that are found to be cost-effective in economic analysis – may not prove the most effective in practice. It is for this reason that health services research has been developed, which seeks to evaluate and inform decisions about the organisation and governance of health services in order to ensure the efficient and effective delivery of the interventions and services accredited in health technology research. The field involves evaluations of service reorganisations and reforms associated with New Public Management, including performance management and marketisation (McPake et al., 2006; Fulop et al., 2003).

At this wider level of governance, RCTs may be difficult to administer because of the considerable complexity of practical health policy environments (McPake et al., 2006). As a response to this challenge, health services research uses observational methods to track the performance of health organisations over the course of a policy change. The preferred method is the prospective observational study which is set up prior to a policy change. The terms of the study are defined prior to the observation, leading to the specification of performance data that is then collected over the certain time period of policy implementation. This provides a ‘before and after’ account of organisational performance and can also be compared with other organisations that have not experienced the change.

Alternatively, retrospective observational studies involve the statistical analysis of existing data sets, often data sets that has been collected for other purposes such as audit or performance management. While these studies can be useful if the time and resources for setting up an experiment or a prospective study are not available, retrospective studies lack the transparency of these other approaches because the terms of the study are not set out before it commences. Retrospective studies can take the form of complex regression analyses and invite the criticism of “data dredging”:

...when subsequent 'hypothesis testing' is based on having seen the results – or worse, on selectively retraining those hypotheses that support a favoured direction (Williams et al., 2002: 106)

Despite this flaw, evidence-based ‘hierarchies’ in health services research consider retrospective observational studies to be superior to qualitative research (Le Grand, 1998). Yet retrospective observational studies have led to significant controversy in health policy debates. The next section explores some criticisms of the methods of Evidence-Based Medicine in terms of their practical implications for health policy.

4. Criticisms of Positivist Approaches to Evaluation in Health

The previous section highlighted the influence of positivism over both health technology research and health services research. The framework is apparent in the construction of evidence “hierarchies” that are used as a guide both for primary research and the appraisal of existing research. Primacy in these hierarchies is given to quantitative methodologies and evaluation has an overarching, positivistic purpose of identifying generalisable solutions to policy problems, whether at the policy level of interventions and services or governance. Lucy Gilson and colleagues argue that the dominance of positivism is such that it is appropriate to talk of “disciplinary capture”, whereby a single research framework has colonised the field (Gilson et al., 2011).
While section 2 suggests that positivist understandings of evaluation and evidence are theoretically problematic, a further sign of its inadequacy is the significant controversies that have accompanied the use of this research in decision-making. Problems have arisen at the level of both policy and governance.

**Evidence-Based Controversies at the Policy Level**

One set of criticisms of Evidence-Based Medicine highlights the influence of pharmaceutical companies over the research agenda. Problems are purported to arise because pharmaceuticals are free to define what constitutes an illness and determine what treatments are tested, the methods used to test them and the outcomes that are measured (Greenhalgh et al, 2014). One of the consequences of this is that more profitable forms of care are developed, discussed in more detail below. A further issue relates to the governance of the research process. Following lobbying campaigns during the 1970s, pharmaceutical companies are not obliged to publish research studies and have a clear incentive to withhold negative findings which present their products in a negative light or to engage in fraud and misconduct when undertaking research (Gupta, 2013).

Somewhat ironically, the claim that pharmaceutical companies distort the evidence-base in health technology research is itself strongly evidence-based. Extensive research has been carried out into this topic and systematic reviews exist which synthesise the research findings. Whether or not the pharmaceutical industry undertakes or funds research is shown to have a significant effect on its quality and/or the availability of research findings (Bekelman et al., 2003; Bourgeois et al., 2010; Heres et al., 2006; Lexchin et al., 2003). One case study of anti-psychotic drugs found that industry-backed publications reported positive results 90% of the time. Many publications reached contradictory conclusions when drugs favoured by one company were assessed by other companies in head-to-head studies7, resulting in an anomaly whereby Drug A (olanzapine) beat Drug B (risperidone), Drug B (risperidone) beat Drug C (quetiapine) and Drug C (quetiapine) beat Drug A (olanzapine) (Heres et al., 2006).

These criticisms highlight issues with the institutional arrangements that determine investment, the development of medical services and ultimately effect the integrity of evaluations. Other criticisms highlight limitations of the scientific method itself. One set of criticisms is aligned with the medical profession and suggests that evidence-based decision-making is not a substitute for professional judgement and expertise. The analysis of RCTs with contradictory results provides evidence to support this view. In a paper entitled ‘Complexity and Contradiction in Clinical Trial Research’, Ralph Horwitz examined 200 RCTs on 36 topics where contradictory results had been reported. These had been put down to methodological deficiencies of the trials or the size of the sample sizes. Yet Horwitz found that many of the contradictory results were not due to methodological deficiencies, but rather the complexity of healthcare itself. Often, slightly different treatments were admitted or different outcomes were analysed. Furthermore, some trials had different eligibility criteria and patients could have different characteristics and states of health. Contradictory results therefore had little to do with methodological rigour but reflected the complex reality of healthcare (Horwitz, 1987).

---

7 The development of medical knowledge has meant that traditional treatments are increasingly outdated. So-called “second-generation” drugs are increasingly tested against traditional treatments and each other in head-to-head studies.
Horwitz’s research implies that clinical guidelines and performance management protocols may still promote the wrong treatments, even when the best evidence is available. This is because each individual case is unique and idiosyncratic. For this reason, evidence must be complemented with professional judgement and expertise, with its “reliance on experience, analogy and extrapolation” (Qizilbash et al., 2008: 5; see also Groopman, 2001; Head, 2005; Meldrum, 2000).8

A still further set of criticisms suggest that scientific forms of evaluation do not provide all interventions with a fair and comprehensive evaluation. This debate mainly concerns the use of pharmacological therapies in preventative care, where drugs such as statins and oral glycaemic agents are used to prevent complications such as heart attacks and diabetes. Proponents argue that preventative care will improve health outcomes and save money in the context of increasing lifestyle conditions and an aging population (Russell, 1993). However, critics argue that medical care has reached a saturation point. The only new therapies are pharmacological therapies which are replacement treatments, with clinical trials evaluating which treatment out of different treatments (as with the case of head-to-head studies), or which combination of treatments, produces minor health gains. Only marginal gains can be expected in the prevention and management of conditions, in what is a “near saturated therapeutic field” (Greenhalgh et al., 2014: 2).

Within the medical profession, there is some scepticism that preventative care is superior to other forms of medical care. A key controversy over preventative medicine is the use of increasingly precise outcome measures in clinical trials. Health interventions tend to be evaluated across hard endpoints and disease-specific effects (see above). Yet preventative care is about preventing any harm from arising in the first-place through the control of biomarkers or surrogate outcomes, which are risk factors for a disease or complications of a disease. The rise of long-term conditions has meant that surrogate outcomes are increasingly central to healthcare: examples include cholesterol levels in heart disease, blood pressure levels in hypertension and blood glucose levels in diabetes.

In preventative medicine, treatments are often evaluated in terms of their effects on surrogate outcomes alone. This is said to be justified where the aetiology of a drug is understood, making it unnecessary for a comprehensive evaluation across hard endpoints. However, improvements across surrogate outcomes do not necessarily translate into improved health outcomes. Some drugs and treatments that improve surrogate outcomes are poorly correlated with actual health outcomes, even detrimental to them (D’Agostino, 2000; Fleming and DeMets, 1996; Psaty et al., 1999). There have been some high profile cases where drugs have been approved on the basis of surrogate evaluations, only to be withdrawn from the market when subsequent research has shown them to be harmful (Hulley et al., 1998; Psaty and Pahor, 2000; Nissen and Wolski, 2007).

Furthermore, pharmacological therapies are typically evaluated and approved for treatment for singular conditions and diseases. But the effects of combinations of drugs are rarely evaluated, either scientifically or economically. In the context of aging populations, where

---

8 In philosophical terms, these arguments resonate with realist criticisms of positivist understandings of science and the distinction between closed and open systems (Bhaskar, 2008).

9 Risk factors are variables associated with a disease but the link is not causal

10 Comprehensive evaluation across hard end-points is more expensive and takes longer than surrogate evaluation, potentially resulting in poor quality healthcare because existing and less effective drugs and treatments would continue to be prescribed (Burzykowski et al., 2006: 4; Cohn, 2004)
comorbidities are increasingly the norm, there is a danger of polypharmacy, whereby individuals are provided with multiple drugs for the different conditions they have. The harmful side-effects of drugs are likely to increase when multiple drugs are used, resulting in health complications and increased cost (Duerden et al., 2013; Munger, 2010). Scientific forms of evaluation therefore appear to overstate the capacity of pharmacological therapies to improve health outcomes, even in a restrictive, medical sense. From a different perspective, others argue they disadvantage holistic interventions which aspire to more qualitative outcomes such as wellbeing and empowerment. There are variants of this criticism in healthcare and public health.

In healthcare, the positive effects of ‘holistic’ interventions, which include exercise and healthy eating, chiropractic care, homeopathy and acupuncture, may be undervalued in scientific evaluation. A key issue is the control of the effects of patient expectations and experiences of care. ‘Blinding’ procedures ensure that patients do not know if they are taking a drug or a placebo (Gensini et al., 2005). Additionally, the effects of doctor-patient relationships and medical environments are controlled, such that outcomes can be attributed to treatments alone. However, these factors have a significant effect on health outcomes. How people perceive their treatments significantly effects how they respond, both psychologically and physiologically (Barrett et al., 2006: 181). It is likely that holistic interventions will be disadvantaged in clinical trials because they require significant input from patients and will therefore only achieve their potential when patients consciously seek them or have a strong relationship with their health professional. Furthermore, the full effects of holistic interventions may not be captured because they are “often indirect and may be intangible”:

This would include, for example, fulfilling one’s potential in life, and increasing personal confidence and self-esteem. Qualitative research with patients who have used some form of CAM suggests that they may experience an array of less tangible benefits, such as a sense of personal empowerment, greater control of one’s condition, enhanced ‘energy’ and opportunities to explore a broader range of ‘causes’ of ill health. Such aspects of health gain are harder to capture through quantitative measures (Hollinghurst et al., 2008: 48)

A similar debate is apparent in public health (Sommer and Parker, 2013: 21). For adherents of ‘medical’ public health, interventions must be clearly evidence-based and have a clear impact on individuals’ health. For adherents of ‘holistic’ public health, in contrast, the turn to science has resulted in a retreat from political issues which affect health, such as poverty, inequality, sexism, racism and homophobia (Fairchild et al., 2010: 58). To tackle these factors requires multiple interventions, but RCTs are unable to accommodate this complexity and ultimately reinforce an individual approach in public health (Baum, 1998).

Evidence-Based Controversies at the Governance Level

Similar criticisms have been made of quantitative evaluations of governance, where there is significant contestation over NPM reforms. NPM typically involves a combination of

---

11 Alternative methods have been developed, called “pragmatic trials”. These combine quantitative and qualitative methods and explore what works, under what circumstances and for different types of people, in accordance with a broad range of criteria (Norquist et al., 1999).

12 According to this perspective, broader forms of public health intervention “stray too far from sober assessment of scientific facts and runs the risk of constituting naked political advocacy”, which ultimately undermines the credibility of the profession and its influence over policy (Goldberg, 2012; see also Rothstein, 2002).
performance management and marketisation. In the health service, New Labour oversaw a significant centralisation of decision-making with health providers performance managed through star ratings and league tables. Furthermore, NICE quality standards and various National Service Frameworks specified exactly what services should be provided at a local level. Some of these Frameworks were linked to economic incentives in order to ensure compliance at a local level. At the same time, New Labour extended previous attempts by the Conservatives to construct a market in healthcare by incorporating a split between purchaser organisations and provider organisations: New Labour’s reforms meant that for the first time private providers were allowed to compete with NHS providers to deliver NHS services.

A plethora of research studies have sought to evaluate these changes. Since no significant reform at the level of governance has been introduced gradually in order that it might be scientifically evaluated, evaluations take the form of retrospective observational studies. This research confronts a number of issues.

In the absence of an experimental evaluation, it is difficult to attribute causation between the change in policy and outcomes. New Labour used a number of policy tools to enhance the performance of health services, including increased spending on health, partnership working, clinical guidance, performance management and competition through the use of markets and the private sector. There is a danger that statistical analyses wrongly attribute an outcome to some policy tool when in fact it has been caused by another. Indeed, this issue of causation was highlighted in fraught discussions over the evidence-base during the passage of the Conservative/Coalition government’s Health and Social Care Act (2011), which went further than New Labour’s reforms, permitting a greater use of markets and the private sector in the health service. The Prime Minister referred to a study published by the London School of Economics in his defence of the reforms:

…competition is one way we can make things work better for patients. This isn’t ideological theory. A study published by the London School of Economics found hospitals in areas with more choice had lower death rates (Greener et al., 2012)

However, critics highlighted significant issues with the study. Cooper et al purportedly demonstrate that New Labour’s market reforms had improved the quality of care of NHS hospitals, indicated by 30 day mortality rates for patients diagnosed with Acute Myocardial Infarction (AMI) (Cooper et al., 2011). However, critics argued that competition is unlikely to have had an impact on the quality indicator because AMI is an emergency procedure where patients are not offered choice13 (Mordoh, 2011: 30). Furthermore, alternative explanations may account for the outcome identified in the study. Pollock et al point to uneven distributions of new technologies, which tend to be concentrated in urban areas where there are more providers (and thus coincidentally more competition), as an explanation for the correlation between competition and improvements in mortality rates (Pollock et al., 2011).

The incident highlights a further issue with retrospective observational studies concerning their lack of transparency. Indeed, critics of the Cooper et al study claimed the authors had engaged in “data dredging”: selecting and manipulating data until a case could be made for marketisation (Pollock et al., 2011). While this may or may not have been the case, in areas where there is significant contestation, as in the case of the health service, retrospective

---

13 Sir Roger Boyle, the then tsar of cardiovascular conditions, called the choice of AMI rates “bizarre”: “AMI is a medical emergency: patients can’t choose where to have their heart attack or where to be treated. It is bizarre to choose a condition where choice by consumer can have virtually no effect” (Greener et al., 2012)
studies are likely to exacerbate the contestation and be rejected or accepted along partisan lines.

A still further issue is the use of quantitative measures to evaluate governance, which is an issue in prospective and retrospective evaluations. As in the case of the quantitative evaluation of medical treatments, the question arises whether all dimensions of performance can be quantified and indeed whether evaluations capture the full spread of costs and benefits following a change in policy. This issue is particularly acute in evaluations of performance management which often rely on the very data collected through performance management.

A number of statistical evaluations of New Labour’s targeted approach to performance management compare the performance of the England NHS (where targets were applied) and the home nations (where targets were not applied) (Alvarez-Rosete et al., 2005; Besley et al., 2009; Hauck and Street, 2007; Propper et al., 2008). Striking differences in waiting times are reported, with sharp declines in England but not in the other nations, which inform the claim that targets are a potentially useful policy tool to enhance the performance of health services (Alvarez-Rosete et al., 2005). Yet this assumes that performance is captured in waiting list data when anecdotal reports suggest that targets resulted in long queues outside hospitals, with patients admitted only when hospitals were confident targets could be met, and ultimately poor quality care due to delays in response times to seriously ill cases (Bevan and Hood, 2006: 530). There is a clear circularity in these attempts to evaluate performance management through an analysis of the data collected through performance management.

Evaluations of clinical guidelines and quality standards face the similar problem that they will be blind to issues at the level of policy regarding the appropriate choice of interventions and services. Indeed, one of the purposes of health technology appraisal is to inform the content of clinical guidance and performance management protocols in order to provide guidance for local actors and to facilitate incentives to ensure compliance. There is always the danger with this approach to clinical governance that professionals are incentivised to deliver inappropriate forms of care. Yet quantitative evaluations do not provide the level of detail required to ascertain whether the guidance and targets have been appropriately defined. ‘Translation Science’ or ‘Implementation Science’ involves the comparison of the activities of health professionals, healthcare organisations and even entire health systems with the best practices embedded in clinical guidelines. A variety of methods are applied in the research in order to identify variations or “evidence-practice gaps” that expose healthcare systems to “unnecessary expenditure resulting in significant opportunity costs” (Grimshaw et al., 2012: 1; see also Grol, 2001). Yet the content of clinical guidance is not evaluated; only whether it has been implemented. Similarly, evaluations of performance management protocols in primary care typically evaluate in terms of whether evidence-based guidelines have been implemented (Calvert et al., 2009; Campbell et al, 2009). This focus on implementation alone is clearly problematic when set aside the limitations of health technology appraisal, discussed above. Furthermore, the implementation of evidence-based guidelines may take place at the expense of other activities that are not covered in the guidelines and which are not attached to incentives (Gillam, 2013). More detailed qualitative research is required in order to ascertain the full impacts of policy and the choices and trade-offs involved in its implementation.

5. Concluding Remarks: The Case for a Qualitative Approach to the Evaluation of Policy and Governance
This paper has criticised the influence of positivism in evaluation in health, exploring a number of alternative research frameworks which suggest that positivism is theoretically problematic and highlighting specific weaknesses with the methodologies used in the field. These weaknesses imply that decisions which rely on evidence alone will be significantly distorted. There is significant concern that the influence of the pharmaceutical industry over clinical research is distorting the evidence-base in health technology research, while excessive centralisation in clinical governance may unduly restrict the scope for the exercise of professional judgement and expertise, an important dimension of which is to interpret and apply evidence to individual cases. Furthermore, dominant quantitative methodologies do not provide all forms of intervention with a fair and comprehensive evaluation. In contemporary medical practice, which emphasises the importance of preventative care in order to prevent diseases and complications from arising, these factors may be combining to contribute to an oversupply of pharmacological therapies when alternative, holistic interventions are more appropriate, in public health and healthcare. A key issue is the use of surrogate outcomes in evaluation in health technology research but also in the commercial marketing of drugs and performance management protocols, where they are increasingly used as indicators of health service outcomes and influence the remuneration of health professionals (Greenhalgh, 1997; Weston, 2008).

Similar issues arise in the quantitative evaluation of governance in health services research. The extent of complexity at the level of governance frustrates efforts to establish causation between policy changes and outcomes, while quantitative evaluations do not provide the level of detail required to capture the full spread of benefits and costs arising from changes in policy. Much of the focus is on implementation alone, but this leaves open the question of the quality of services and forms of care that are actually produced. Without addressing this question, evaluations in health services research provide a skewed account of the performance of health services in general. Indeed, Kabir Sheikh and colleagues criticise the focus on implementation in the field, calling it the evaluation of “predetermined programmatic solutions” (Sheikh et al., 2011: 4; see also Gilson et al., 2011).

This is not to argue that quantitative evaluations in health technology research and health services research have no value, but there is a need to complement them with qualitative research. To that end, frame analysis might be adapted for the purposes of policy and governance evaluation. Frame analysis recognises the potential for multiple interpretations of reality, value diversity and the fragmented nature of knowledge. Frames organise experience and provide something of a narrative that interprets the world, determining what is sayable. They are relatively durable, are grounded “in the institutions that sponsor them” (Schon and Rein, 1995: 29) and serve a political function, seeking to elicit particular responses from actors (Payne, 2001: 29). They contain both diagnostic and prognostic elements, identifying what is wrong and how problems might be solved (Lombardo et al., 2009: 141).

Exploring stakeholders’ framings in empirical research might reveal significant issues at the levels of policy or governance. A key feature of frames is the values which underpin them. Indeed, where stakeholders are asked to evaluate policy and governance, their evaluations will reveal certain values and criteria. It is important to consider such values in order to ascertain the level of contestation at stake in any dispute. It may be the case that stakeholders identify the suppression of certain values that are not recognised by others, suggesting the presence of significant political contestation. Where such a conflict exists, the role of policy analysis can be to elucidate the nature of the dispute such that it can be discussed forthrightly in the political domain. Still, there may also be occasions where shared
values are present, suggesting that any stakeholder’s knowledge of a detrimental policy impact would be a mutual interest.

A further key feature of frames is the knowledge they include. Stakeholders’ evaluations of policy and governance might draw attention to issues at the level of policy pertaining to the nature of the issue at hand and the interventions and services that are (or potentially could be) used to solve it. Given the increased importance of Evidence-Based Medicine at this level, moreover, stakeholders might put forward their own interpretation of the evidence-base, highlighting weaknesses which other stakeholders are not aware of. Moving beyond this level of policy, stakeholders might draw attention to issues with governance. Indeed, it is important to consider both levels because the effectiveness of each depends to a large extent on the effectiveness of the other. Stakeholders might highlight issues with central or local decision-making or questionable effects of the incentives-environment, again which other stakeholders are not aware of. Where issues of this sort are identified, the role of policy analysis can be to bring attention to the issue and identify possible solutions if at all possible. This approach is applied in a case study of diabetes in a forthcoming paper.
Bibliography


Greener, I., Macfarlane, A., Pollock, A.M., 2012. Bad science concerning NHS competition is being used to support the controversial Health and Social Care Bill. Lond. Sch. Econ.


