The Quality and Outcomes Framework (QOF) and General Practice Quality in England[*]

Click here for the latest version.

Theodore Caputi

March 23, 2020

Contents

[*] These results are preliminary.

List of Figures

List of Tables

1 Introduction

Quality in health services has been the central concern of England's National Health Service (NHS) since the 1998 publication of *A First Class Service* (Secretary of State for Health, 1998). *A First Class Service* formally placed the responsibility of improving quality on all NHS organisations and employees and laid out a new system of clinical governance, defined as "a framework through which NHS organisations are accountable for continuously improving the quality of their services and safeguarding high standards of care by creating an environment in which excellence in clinical care will flourish" (Secretary of State for Health, 1998), through which NHS service providers would be expected to continuously improve quality. Due to the centrality of general practice in England, improving quality in general practice immediately became a top priority for the NHS (Baker et al., 1999).

The Qualities and Outcomes Framework (QOF) is the latest and most comprehensive in a series of schemes adopted by the NHS to improve quality among general practitioners (GPs). Overall, the QOF is a voluntary but widely adopted pay-for-performance scheme in which GP practices are rewarded for scoring points through delivering a set of predetermined interventions and reaching a set of predetermined targets. The QOF scheme was first introduced to GPs in 2004 as part of the General Medical Services contract between the NHS and individual GP practices. At the time of its adoption, the QOF included 146 indicators within four domains: clinical, organisational, patient experience, and additional services (Downing et al., 2007). Practices could earn up to 1050 total QOF points depending upon their level of achievement on each of the indicators, with certain indicators weighted more heavily than others. QOF points were assessed by practice each year using the Quality Management and Analysis System, and GP practices were awareded a financial reward based upon their achievement. The QOF was initially hailed as both offering "the promise of a quantum change in performance rather than an incremental one" (Shekelle, 2003) as well as "medicine by numbers... threaten[ing] the professional basis of general practice" (Lipman et al., 2005).

Since its adoption, the QOF scheme has undergone several changes. For example, an expert panel involving participants from the University of Birmingham, the Society for Academic Primary Care and the Royal College of General Practitioners were appointed in 2005 to review the QOF indicators, which led to the addition of seven new clinical areas beginning in 2006 (Lester et al., 2006). Since 2009, the National Institute for Health and Care Excellence, the UK's governmental body responsible for facilitating the translation of research into practice, has been tasked with leading the process of setting QOF indicators (Sutcliffe et al., 2012). As of 2018/2019, the QOF includes three domains (Clinical, Public Health, and Public Health – Additional Services) involving 65, seven, and five indicators, respectively (National Health Service, 2018*a*). Now nearly 15 years since its initial implementation and with adoption rates above 94% (participation of 94.8% in 2017-2018; (National Health Service, 2018*b*)), the QOF scheme is a well ingrained component of general practice in the United Kingdom.

1.1 History of General Practice Quality

1.1.1 The Collings Report Reveals Shortcomings in Patient Experience

Quality in General Practice has been a concern in the UK since the formation of the National Healht Service (NHS). When the NHS was founded in England in 1948, just three years after the end of the devastating World War II, general practice became formally enshrined as the core of healthcare in the UK, responsible for all personal healthcare and acting as a gateway to specialised and hospital services (Goodwin et al., 2011). It did not take long thereafter for GPs to garner criticism. In 1950, Dr. Joseph S. Collings (Collings, 1950) published a long-form (30 page), ethnographic study in *The Lancet* entitled "General Practice in England Today - A Reconnaisance". Collings had experience as a physician in Canada and New Zealand and was a research fellow at Harvard University's School of Public Health (Petchey, 1995), but before conducting his study, had never visited the UK. In his study, Collings described his observations of day-to-day GP activity from 55 English GP practices that he had visited, representing industrial, urban-residential, and rural area practices. He concluded that "the overall state of general practice is bad and still deteriorating" and that the conditions of English GP varied markedly across industrial, urban-residential, and rural area practices. He described industrial practices as inhospitable, overcrowded, and untidy and remarked that consultations at industrial practices were cursory and incomplete. While he described rural practices as "an anachronism", he found they were relatively better than industrial practices, applauding their safety and comfort and the organisational prowess of rural doctors. Urban-residential practices spanned the difference between industrial and rural area GPs. Collings vehemently dismissed the notion that quality would inherently be lower in industrial rather than rural practices, insisting instead that "[i]f the doctors' surgeries were better from a functional point of view; if they were equipped at least to some minimum standard; if a little order were brought into the chaos which characterises the organization of the average practice (e.g., by employing clerical assistants); and if the work of the industrial practitioner were coordinated with that of other medical agencies-then the same volume of work could be handled at a much higher level and with due regard to the safety and comfort of the patient" (Collings, 1950). Collings described that the NHS had not planned for or incentivized good practice and that professional and administrative checks were insufficient. Ultimately, he recommended that a system be developed in which GP practices operate out of state-run facilities (Fry, 1988).

While Collings' report was methodologically limited (many criticized Colling's sampling strategy and his observational methods (Petchey, 1995)), its rich descriptions resonated strongly with contemporary GPs (Honigsbaum, 1979). It spurred responses from academics and physicians within (Hunt, 1955) and outside of *The Lancet* (Anthony, 1950; Dawson, 1950; Lask, 1950; Pirrie, 1950; Sanguinetti, 1950) and, by extension, sparked widespread policy interest in the evaluation and improvement of quality among NHS GPs (Petchey, 1995). For example, the Colling's report is specifically cited as the impetus for the 1952 creation of the College of General Practitioners (now the Royal College of General Practitioners) (Royal

College of General Practitioners, 2019). As stated on the College's website, the College of General Practitioners was founded under the "shared belief that what was needed [to address the issues raised in the Collings report] was an academic body to support good standards of practice, education and research, such as already existed in other medical colleges" (Royal College of General Practitioners, 2019).

1.1.2 The Hadfield Report Shifts Attention to Burden of GPs

Many scholars suggest that the 1950 Collings Report directly prompted another major investigation of GPs in the UK: the 1953 Hadfield report (Wilkie, 2014). The Hadfield Report (Hadfield, 1953), sponsored by the British Medical Association's General Practice Review Committee and authored by Committee Secretary Dr. Stephen J. Hadfield[1], surveyed the experience of 200 GPs located in England, Scotland, and Wales between February 1951 and March 1952. Seemingly to improve upon Colling's sampling strategy, GPs were selected near-randomly from 20 separate regions (the 19 regional hospital board areas and the Metropolitan Branch of the British Medical Association). Hadfield visited each practice, sat in on at least one consultation, and then accomopanied the GP on his normal rounds. He documented his impressions on the lifestyle of the "average" GP (e.g., where they live, their daily schedule, their leisure time), the services that GPs provide (e.g., continuity of treatment, preventive medicine, diagnosis, treatment, providing information to specialists, non-medical advice, clinical assistantships in hospitals, and administrative duties), the professoinal relationships GPs hold (e.g., with other GPs, with consultants, with public health officers, with nurses), the organisation of GP practices (e.g., premises, equipment, appointment management, employment of non-medical staff, and record-keeping), and views that GPs hold on auxillary services (e.g., effect of the NHS on general practice, services provided to health visitors, the adequacy of hospital services, and the utility of health centers). Hadfield described several shortcomings of the GP system throughout his report and painted

[1] It should be noted that the British Medical Association strongly supported the newly implemented National Health Service (Anonymous, 1953).

general practice in the UK as underfunded and demoralising work. However, he ultimately finds that 44% of practices had overall “good” quality, 44% had “adequate” quality, and only 7% had “inadequate” quality (5% were not assessed).

Hadfield concluded his report with three “adverse impressions” from his investigation. First, he criticized the lack of cohesion and continuity within the NHS, including weak relationships among GPs, specialists, and public health medical officers and discontinuities in administration. Second, he criticized inadequate and lagging hospital services, which place additional strain on GPs. Finally, he criticized GPs response to an increased demand for services. He described that the public’s expectations of entitlements from the NHS are already “deep-rooted” and argued that, given the increased responsibility of GPs and the increased expectations from patients, a strategy to handle increased demand was necessary. Ultimately, the Hadfield Report, while considered more optimistic than the Collings Report, affirmed that there was substantial room for improvement in UK general practice. An anonymous editorial appearing alongside the Hadfield Report in *British Medical Journal* stated: “[i]t is a picture that justifies a restrained optimism for the future of general practice – restrained because, unless remedies are applied soon and continuously, the present good service the public receives from general practitioners will deteriorate in time”.

To some degree, it appears that the legacy of the Hadfield Report was shifting public attention away from patient experiences, the focus of the Colling’s Report, and towards the frustrations faced by GPs under the new NHS regime. For example, soon after the publication of the Hadfield Report, Dr. Stephen Taylor, a doctor and former Labour politician, authored a report entitled “Good General Practice” recommended that GPs could improve quality by reducing their personal burden (Taylor, 1954). Taylor concluded, ”good general practice begins with the good GP... [s]o most of the conclusions are suggestions for self help” (Taylor, 1954).

1.1.3 Redressing Greivances with the NHS

By 1965, GPs had accumulated several grievances with their position within the NHS. GPs argued that the overwhelming demands of the job, combined with no formal provisions for quality improvement, left them incapable of providing good quality service to patients. The British Medical Association described these grievances in its "Charter for the Family Doctor Service", which set forth what GPs needed in order to offer high-quality service to their patients. Published in *British Medical Journal*, the Charter demanded a lightened burden so that GPs could properly practice medicine (GMS Committee, 1965):

> If the trained doctor who enters general practice is to be able to do those things he has been trained to do, then he must have adequate time, space, and assistance with which to do them. In order to have these basic requirements he must find money to provide the space, equip it properly, and engage the services of non-medical helpers who in their own special field of experience can take a good deal of work off his hands, and so leave him free to do that which he has been trained to do. Above all, the general practitioner must have time to take a full history and make a careful examination of those patients who come to him with other than trivial ailments, in the management of which many were content to take the advice of the local pharmacist. The characteristic of modern medicine is increasing precision of measurement. This is the characteristic of science. The general practitioner should therefore have fully available to him those facilities provided by pathology and radiology which enable him to fill out the clinical picture and to confirm his suspicions or guesses, and in any case to provide precise data which would clinch a diagnosis and facilitate rational therapeutics. Many general practitioners work to-day in conditions which make this difficult, or even at times impossible. A shrewd clinical instinct developed by experience and guided by intuition often makes good to a surprising extent these inadequacies. But the modern patient and the modern doctor both demand the full use of the resources of modern clinical science. These things, though commonplace enough, have to be repeated if the public and the Government are to realize that doctors are not just crying for the moon-or even for the moon and sixpence. Medicine advances at such a rate that unless a doctor once qualified deliberately sets aside time for reading, and periodically time for retraining, he is bound to get out of date and to feel himself isolated from his colleagues who, by continuing to work in hospital, are confronted throughout the whole of their work with new ideas and with frequent interchange of these and of experience with other hospital colleagues. This continuing education needs time, and time is money.

The following year, the demands of the Charter were accepted and codified in the 1966 NHS General Practice Contract, which massively improved GP pay and conditions (Gillam, 2017). For example, the Contract reduced capitation-based income and increased practice allowances and fee-for-service opportunities, increasing overall pay to GPs. In addition, the Contract offered reimbursements for staff-related expenses, allowing GPs to hire nurses and non-clinical staff, and subsidies for facility development. The Contract further placed a limit on a GP's patient list size at 2,000, reducing the maximum number of patients a GP would be responsible for. The advantageous provisions of the Contract attracted a new wave of people to become GPs (Horder and Swift, 1979), and the era immediately following 1966 has been considered a "golden age for general practice" (Addicott and Ham, 2014).

1.1.4 Internal Regulation of Quality

GP Associations had demonstrated their political dominance over the Government and the NHS in the 1966 General Practice Contract negotiations, and in subsequent decades, healthcare quality was monitored, standardized, and addressed internally through professional medical organisations (Howie, Heaney and Maxwell, 2004). The College of General Practitioners had worked to establish standards of practice, offering "Foundation Membership" to GPs who satisfied a defined set of criteria (Royal College of General Practitioners, 2019). In 1972, the College of General Practitioners was awarded a formal Royal Charter, which established the organisation (renamed the Royal College of General Practitioners) into the first official professional organisation for GPs. Four years later, the Royal College of General Practitioners successfully lobbied Parliament to approve the 1976 NHS Vocational Training Bill, which required three years of vocational training to become a GP and established national standards organisations to monitor that training. In turn, the number of GPs enrolling in formal vocational training to become a GP skyrocketed, increasing 130% between 1976 and 1985 (Petchey, Williams and Baker, 1997).

By the 1980s, external researchers had accumulated substantial evidence that the quality of general practice in the UK was inconsistent and that certain groups of patients were receiving inadequate care (Baker, 1989). The public, who had increased oversight of other government agencies, grew discontent. To mitigate some of this criticism and to reduce the probability of a government-led reform, the Royal College of General Practitioners launched its "Quality Initiative" in 1983, which sought to improve professional development of doctors, practice management and team-work, quality assessment, contracts and incentives, and resource needs (Practice, 1985). Generally, the Initiative was intended to raise awareness about quality measures among GPs and to encourage GPs to consider how they could improve the quality of their own practice (Baker, 1989). Specific recommendations from the "Quality Initiative" included the following (British Medical Journal, 1985):

- Higher training through appropriately supervised experience in the early years of general practice should now be introduced to promote the further professional development of young principals.
- The concept of "protected time" for learning should be extended to all principals and practice staff and the resources necessary for this should be made available.
- A credible entry standard for NHS general practice should be determined. In future, doctors entering the NHS list as principals should be encouraged to reach the standard available through the membership of the Royal College of General Practitioners examination.
- The completion of higher training should be denoted by accreditation given by the college, and should be based on the continuous assessment of clinical performance as in other specialties of medicine.
- Good quality practice needs good quality management. Every practice should provide the basic range and quality of services which every patient can expect anywhere.
- Standard setting and performance review are activities that must be incorporated into everyday clinical practice. They make essential contributions to effective practice management and to the continuing education of members of the practice team.
- Incentives should be developed to encourage doctors to participate in performance review. A solution should be found within the framework of the National Health Service to bridge the gap between how doctors are paid and the standard of services that they provide. Unacceptable levels of performance should be reflected in a doctor's remuneration.
- Resources must be found for the implementation of the RCGP's council's strategy for quality.
- Membership of the college should be open to older doctors who submit successfully to a system of assessment that is no less rigorous than the MRCGP examination.
- Performance review should be incorporated into every member's professional life, and such participation should contribute to the attainment of the fellowship of the college.

1.1.5 The Managerial Revolution and Market-Based Mechanisms

The light-touch quality improvement strategy implemented by the Royal College of General Practice was well-received by GPs (Baker, 1989) but failed to ward off criticism. Inspired by the Griffith Report of 1983, policy-makers began calling for the government to take on tigheter managerial responsibility over GPs and to apply the general management tools from private industry into healthcare (Strong and Robinson, 1990). For the first time, the Government began discussing market-based interventions designed to incentivize high-quality treatment. In 1986, the Secretaries of State for Social Services released a "Green Paper" that recommended a pay-for-performance system called the "Good Practice Allowance", which would provide financial rewards to GP practices that passed an inspection and achieved a certain rate of screening and preventive services (Baker, 1989). There was sufficient backlash from the medical community for the government to remove all mention of the "Good Practice Allowance" in its follow-up 1987 "White Paper", but other market-based incentive mechanisms persisted. For example, the Government recommended that GP practices be encouraged to advertise their services, increasing competition among practices, and that GP practices would be given some financial compensation for increasing their list size. A small set of preventive services would also garner an additional fee.

These new managerial systems were formally adopted in the 1990 NHS General Pracitoner Contract (Williams et al., 1993). By increasing the GP's variable proportion of income (e.g., capitation-based and fee-for-service payments), the new contract was designed such that funding would "follow the patient", meaning that GP practices (as well as acute care, community care, priority care, and health promotion units) were financially incentivized to compete for patients. Importantly, the 1990 Contract included additional incentives for meeting or exceeding certain "performance related" or quality benchmarks, such as immunisation rates. In this way, GPs were, for the first time, financially incentivized to achieve quality standards.

[Placeholder for Paragraph on GP Fundholding]

A further step towards UK Government monitoring of healthcare quality was the launch of the National Institute for Clinical Excellence (NICE) on 1 April 1999. NICE was formed with the understanding that physicians were unable to stay up-to-date with best practices for care and was tasked with providing physicians and other healthcare professionals in the NHS (including GPs) with the correct tools to bridge clinical care and best practice (Rawlins, 1999). Specifically, NICE was given three main responsibilities: (A) appraising new health technologies, (B) developing clinical guidelines from the literature, and (C) promoting clinical audits and inquiries. NICE's actions on clinical guidelines and audits laid the groundwork for an expanded presence of Government in quality monitoring. The clinical guidelines set standards which could be adapted into performance indicators (Campbell, 2002), and NICE's new oversight of clinical inquiries moved the balance of power away from the Royal Colleges, who remained responsible for conducting National Confidential Enquiries.

1.1.6 The Current Quality and Outcomes Framework Scheme

By the early 2000s, the field of general practice was in turmoil, facing decreases in both GP morale and in new GP recruitment (Roland and Campbell, 2014; Roland and Guthrie, 2016). The NHS worried that these factors could lead to decreases in quality. To ameliorate the situation, the Government offered GPs a significant pay raise in return for greater accountability for the quality of their services. The GPs agreed. With the 2004 General Medical Services contract between the NHS and individual GP practices, the NHS's initial forrays into pay-for-performance grew into the dramatically expanded Qualities and Outcomes Framework (QOF) scheme, and average GP pay was increased by nearly 20%. The QOF was controversial among practitioners and policy-makers. Some described it as "the promise of a quantum change in performance rather than an incremental one" (Shekelle, 2003), while others criticized it as "medicine by numbers... threaten[ing] the professional basis of general practice" (Lipman et al., 2005).

The QOF was and remains a voluntary but widely adopted pay-for-performance scheme in which GPs are awarded payment-linked points for delivering certain interventions and achieving certain quality targets. The QOF initially included 146 indicators within four domains: clinical, organisational, patient experience, and additional services (Downing et al., 2007). Practices could earn up to 1050 total QOF points depending upon their level of achievement on each of the indicators, with certain indicators weighted more heavily than others. QOF points were assessed by practice each year using the Quality Management and Analysis System, and GP practices were awareded a financial reward based upon their performance, with adjustments made for workload and the health of patients in the practice's area.

The QOF scheme has changed several times since its adoption. For example, in 2005, an expert panel involving participants from the University of Birmingham, the Society for Academic Primary Care and the Royal College of General Practitioners were appointed in 2005 to review the QOF indicators, which led to the addition of seven new clinical areas beginning in 2006 (Lester et al., 2006). Since 2009, the National Institute for Health and Care Excellence, the UK's governmental body responsible for facilitating the translation of research into practice, has been tasked with creating a menu of "suitable" indicators that could be included in the QOF scheme (Sutcliffe et al., 2012). As of 2018/2019, the QOF includes three domains (Clinical, Public Health, and Public Health – Additional Services) involving 65, seven, and five indicators, respectively (National Health Service, 2018*a*). Now nearly 15 years since its initial implementation and with adoption rates above 94% (participation of 94.8% in 2017-2018; (National Health Service, 2018*b*)), the QOF scheme is a well ingrained component of general practice in the United Kingdom.

1.2 Frameworks and Theories

1.2.1 Definitions and Frameworks of Healthcare Quality

Quality is a famously nebulous term. Even within the context of healthcare, various conceptions of quality exist. Consequently, when evaluating quality, it is useful to be aware of the various definitions and frameworks that may be applied.

Definitions of quality vary substantially and emphasize different points. For example, (Donabedian, 1980) focused on the benefit-risk ratio, defining quality as “the application of medical science and technology in a manner that maximises its benefit to health without correspondingly increasing the risk”. Ovretveit and Townsend (1992) extends the endgoal of quality from simply maximizing benefit and minimizing risk to also exceeding patient expectations: “Provision of care that exceeds patient expectations and achieves the highest possible clinical outcomes with the resources available”. Schuster, McGlynn and Brook (1998) define good quality care as “providing patients with appropriate services in a technically competent manner, with good communication, shared decision making, and cultural sensitivity”, matching thier framework. Lohr (1990) emphasizes the role of professional standards in achieving quality, defining quality as “the degree to which healthcare services for individuals and population increases the likelihood of desired health outcomes and is consistent with the current professional knowledge”, a definition formally endorsed by the Institute of Medicine (Schuster, McGlynn and Brook, 1998).

Most frameworks for understanding health quality attempt to break down the vague concept of “total quality” into domains or perspectives that evaluators can explore separately (Mosadeghrad, 2012). However, these domains or perspectives vary significantly. For example, the framework put forth by Ovretveit and Townsend (1992) considers three different dimensions: professional quality, referring to the use of the correct procedures to meet patient need; client quality, referring to satisfaction from the patient; and management quality, referring to efficiency of healthcare delivery. Joss and Kogan (1995) consider quality in three different domains: technical quality, referring to the procedures performed; systemic quality, referring to the overarching systems and processes that the procedures fit into; and generic quality, referring to interpersonal relationships. Schuster, McGlynn and Brook (1998) focus on providing the appropriate level of care and assess shortcomings in terms of too little care, too much care, or the wrong care.

Perhaps the best known conceptual model for evaluating quality in healthcare was developed in 1966 by Dr. Avedis Donabedian of the University of Michigan on commission from the Health Services Research Study Section of the United States Public Health Service (Donabedian, 1966). Donabedian conceived of healthcare as a dynamic system that could be evaluated in terms of its processes, structures, and outcomes. He felt that quality was ultimately determined by important patient outcomes but recognized that health outcomes were exceedingly difficult to accurately observe. His framework, consequently, combines his focus on outcomes with detailed observation of structures and processes that are closely linked with clinical outcomes. When an evaluator uses the Donabedian framework to evaluate healthcare quality, she observes the structure of healthcare, i.e., the context in which medical care is delivered; the process of healthcare, i.e., the actual delivery of care; and outcomes, i.e., the effect of the healthcare delivered on the morbidity or mortality of an individual or population. Donabedian's conceptual model persists as perhaps the most common framework for evaluating healthcare quality (Berwick and Fox, 2016).

Aggregating definitions of and frameworks for quality by past scholars, Mohammad (2013) develops a particularly useful meta-framework. In this meta-framework, he distinguishes between quality evaluated from the perspective of healthcare services suppliers and quality evaluated from the perspective of healthcare services demanders. On the supply-side, quality can be assessed in relation to a set of benchmarks predetermined by managers. Benchmarks are set, *before* any service is renderred, in relation to what past experience (e.g., research literature or expert opinion) would dictate is best practice. The definition for quality set forth by Lohr (1990) would fit into this category. On the demand-side, however, quality is measured in relation to patient-specific needs and expectations. In this way, quality is assessed as a matter of satisfaction or dissatisfaction by the patient. The definition of quality set forth by Ovretveit and Townsend (1992) reaches into this domain.

1.2.2 Theoretical Approaches to Quality Improvement in Healthcare

Over the past 70 years, scholars and policy-makers have debated the merits of various approaches to maintaining and improving quality in healthcare. The history of general practice in the UK has been dominated by four broad models, commonly referred to as the "Altruism", "Hierarchy and Targets", "Reputation", and "Choice and Competition" models (Bevan and Fasolo, 2013).

The Altruism model, also known as the Trust model (Le Grand, 2007), is based on the notion that people will try their best and achieve the best possible outcomes given their resources. In that vein, the Altruism model suggests that the quality of healthcare will improve if GPs are given better resrouces and information. According to the Altruism model, it is mostly useless to either punish or incentivize good or bad performance, as any observed variation in performance is externally determined and cannot be changed by the GP. Instead, under this model, poor performance should be met with increased investment to improve those GP's resources and information (Powell, 1976). For the first several decades of the NHS's existence, this was a particularly compelling model for physicians, including GPs (Bevan, 2010). Medicine was one of the most highly-trusted professions, and people valued not just the care they received from their GP but the personal relationship they had with their GP. The public felt uncomfortable with questioning the motivation of physicians. The Altruism model is low-cost and amenable to GPs (Le Grand, 2007), but it requires the strong assumption that all GPs are exerting maximum effort.

The Hierarchy and Targets model (also referred to as the Mistrust model (Le Grand, 2007)), on the other hand, accounts for variation in GPs' motivation through external incentives. This model relates closely to the theoretical economic literature on principals and agents (Jensen and Meckling, 1976). The principal – in this case, the Government or the taxpayers – pays the agent – in this case, GPs or GP practices – to perform some work. With perfect information, the principal can offer to pay the agent only if she performs at her maximum capacity. However, the Hierarchy and Targets model assumes that there is asymmetric information; that is, the principal can only observe the work the agent completes but not the full-effort capacity of the agent. Conceptually, this asymmetrical information benefits the agent, who can shirk (reducing the agent's disutiliy of effort) without the principal knowing whether the limiting factor on output is the agent's capacity or her level of effort. If the principal wants to the agent to maximize effort and output, then, he has to provide the agent with incentives to the output (Conn, 1982). Consequently, the Hierarchy and Targets model calls for financial incentives (or disincentives) on the quality of GP's services. The principal sets certain standards for the agent's output, and the agent is rewarded based upon the extent to which her work meets those standards. This model underlies pay-for-performance schemes, through which GPs are rewarded if they meet certain preset targets for quality. The Hierarchy and Targets model is intuitively appealing but also comes with high monitoring costs. Further, financial incentives for high-quality tend to be unpopular among physicians, who lose their ability to costlessly shirk (Le Grand, 2007).

In some ways, the Reputation model combines aspects of the Altruism model and the Hierarchy and Targets model. According to the Reputation model, agents can be incentivized to maximize effort not only by financial incentives but also by threats to their reputation. This model assumes that physicians enjoy their status as being highly trusted, which, in turn, means that they'd be willing to exert some additional effort to protect their reputations (Wachter, 2013). Reputation models are often implemented by publicizing the quality of care for individual GPs or practices (Jarvis, 2012). For example, the NHS follows the Reputation model when it shares individual practice's QOF scores online. Ideally, information shared through this model will be widely disseminated and easily understood, such that exerting less than maximum effort has a meaningful reputational impact for the GP. Similarly to the Hierarchy and Targets model, the Reputation model is intuitive but also costly and unpopular among GPs (Le Grand, 2007).

The Choice and Competition or Quasi-Markets (Le Grand, 2007) model is similar to the Hierarchy and Targets model in that it offers financial incentives (or disincentives) to GPs for exerting high effort (or low effort). However, instead of giving principals the responsibility of determining standards of quality, the Choice and Competition model allows consumers to decide what constitutes high quality service. This model assumes that patients – healthcare consumers – are sophisticated arbiters of quality and allows them to choose from a market of healthcare providers. The providers are, in turn, rewarded based upon the number of patients they serve and the extent of services they provide. In sum, GPs exert effort to maximize quality, so that they will attract patients, which, in turn, increases their payments. In a Choice and Competition model, information about GP quality is made publicly available (and ideally, easily accessible) and patients are given the autonomy to choose their providers. The Choice and Competition model is theoretically appealing as it increases patient autonomy, avoids the issue that Governments may choose suboptimal quality standards, and has low monitoring costs. However, it also increases the search costs on patients and relies on the strong assumption that individuals will be (or become) good arbiters of quality. While there is some evidence that patients' healthcare decisions respond to quality (Howard, 2006), these transaction costs may be substantial (Magee, Davis and Coulter, 2003). If these transaction costs are too burdensome for many patients, who then choose suboptimally (e.g., prefer to remain at their original practice), the incentive for GPs to improve quality may be lost (Fotaki, 2013).

Each of these models has played a role at some point in the history of general practice within the NHS. The Altruism model matches early policies intended to improve quality. The Hadfield Report (Hadfield, 1953) and later the Charter for the Family Doctor Service (GMS Committee, 1965) argued that quality was low because doctors did not have the resources to adequately practice medicine. In turn, the 1966 NHS General Practice Contract was signed to increase GP pay and conditions (Gillam, 2017), with the understanding that increased resources would allow GPs to improve quality. Internal regulation of quality by bodies such as the Royal College of General Practitioners could be considered a form of the Reputation model. While there were no financial incentives or penalties, these professional organisations created a reputational incentive for adhering to certain standards by setting quality requirements for membership (Royal College of General Practitioners, 2019). When the public became wary that internal regulation may be too lenient, the Government called for market-based incentives. In line with the Choice and Competition model, the Government suggested that funding should "follow the patient", incentivising practices to compete among themselves (Baker, 1989). These suggestions became part of the 1990 NHS General Practitioners Contract, along with some additional incentives for achieving certain targets, in line with the Hierarchy and Targets model (Williams et al., 1993). Since the early 2000s, the Government has invested heavily in the Hierarchy and Targets model with the adoption of the Quality and Outcomes Framework in 2004. QOF, the largest pay-for-performance scheme in the world, sets certain quality targets and provides financial rewards to practices that meet or exceed those targets, giving GPs a financial incentive to achieve quality standards (Shekelle, 2003). The Hierarchy and Targets model remains the dominant model for quality improvement in the UK.

1.2.3 Challenges to Measuring Quality in Healthcare

There is a substantial literature documenting the challenges involved in evaluating quality in service industries (Proctor and Wright, 1998), and healthcare quality is no different (Mosadeghrad, 2014*a*,*b*; Cheng, Tang and Jackson, 1999). Frequently cited challenges to assessing quality in the healthcare industry include patient participation, simultaneity, perishability, intangibility, and heterogeneity (Ozcan, 2005).

- **Patient Participation**: Healthcare delivery involves highly complex and dynamic interactions between patients and an array of healthcare professionals (e.g., physicians, nurses, pharmaceutical developers, pharmacists, public health officials, health product producers), and so it can be difficult to trace a causal pathway from any one actor's actions and patient outcomes. If a patient resists the best practice care offered to them or chooses to ignore a physician's advice and suffers worse clinical outcomes, should that reflect poorly on the healthcare system?

- **Simultaneity**: Healthcare is simultaneously produced and consumed, making it challenging for quality control protocols to proactively promote good quality care. In the case of manufactured goods, a quality control officer could pick goods that fail to meet certain standards from the shelf; this is not possible with healthcare services.

- **Perishability**: Healthcare services are largely perishable. For example, physicians typically provide services during a prescribed set of hours – if some fraction of those hours are unoccupied, the capacity of that physician is essentially wasted. Given the unpredictable nature of health, it may be impossible for physicians to accurately match the timing of patient demands for healthcare with their availability.

- **Intangibility**: As an intangible "good", health is subjective, multifaceted, and fleeting, which makes outcomes in health more difficult to quantify than physical goods.

- **Heterogeneity**: Heterogeneity in the needs of the patient, even within certain well-defined conditions, makes it challenging to compare services rendered with uniform benchmarks. The practice of medicine requires countless "judgement calls", in which the physician has to make decisions based upon nuances that could not be prescribed in professional guidelines.

It is potentially impossible to successfully disentangle several of these complications from any system of measuring healthcare quality, particularly given resource constraints on healthcare quality monitoring. Instead, healthcare quality monitors have largely accepted these limitations and focused on selecting and applying sets of performance indicators in ways that mitigate these concerns as best as possible (Bennett et al., 2014).

1.2.4 Best Practices for Selecting Quality Indicators

By the turn of the century, as institutions were becoming increasingly concerned with healthcare quality and performance indicators were becoming more fundamental to the healthcare system in the UK and internationally, there already existed decades worth of small-scale and anecdotal research on how managers chose and used healthcare quality indicators (Starkweather, Gelwicks and Newcomer, 1975; Grimes and Moseley, 1976). Realizing that quality indicators were to become more systematically important to the healthcare system, scholars sought to synthesize the findings from these past experiences and inform policy-makers of the best practices for selecting performance indicators. Freeman (2002) conducted a systematic review of the literature on performance indicators. He found that summative performance indicators designed to reward or punish individuals tended to corrupt the indicators themselves, presumably because they failed to account for heterogeneity and the number of "judgement calls" in clinical medicine. On the other hand, he found that formative indicators gave clinicians clues of how they could improve while allowing clinicians and their managers to make adjustments for local and clinical context. (Mainz, 2003*b*) used the experience of the Danish National Indicator Project to outline a specific process for developing the right indicators. He recommended that clinical indicators first be defined and prioritised based upon the scientific literature, then specific targets should be adopted with special consideration given to each indicator's measurement and inclusion/exclusion criteria. He also characterized the features of an ideal healthcare indicator (Mainz, 2003*a*):

- Indicator is based on agreed definitions, and described exhaustively and exclusively
- Indicator is highly or optimally specific and sensitive, i.e. it detects few false positives and false negatives
- Indicator is valid and reliable
- Indicator discriminates well
- Indicator relates to clearly identifiable events for the user (e.g. if meant for clinical providers, it is relevant to clinical practice)
- Indicator permits useful comparisons
- Indicator is evidence-based

The concepts behind these criteria were and remained highly influential. For example, these same themes can be observed in the criteria for ideal indicators provided by Campbell and Lester (2010) almost a decade later.

- Acceptibility: Both GPs and patients will accept assessment of the indicator.
- Attributable: A GP's performance on an indicator relates well to his or her own performance and not to external factors.
- Feasibility: Reliable data is available.
- Reliability: There is minimal measurement error in the indicator.
- Sensitivity to Change: Changes to clinical practice can be observed through the indicator.
- Predictive Value: The indicator is closely linked with important clinical Outcomes.
- Relevance: There is evidence of incongruity between actual practice and best practice.

Building upon these research studies, governmental bodies and medical organisations weighed in on how to best choose performance indicators. The Royal Statistical Society's Working Party on Performance Monitoring the Public Services published a report in 2004 that included a list of fourteen guidelines for appropriate indicators in evaluating the quality of healthcare (as well as other government services) (Bird et al., 2005). Among other recommendations, the Working Group suggested that indicators should be directly linked with the NHS's primary objectives, assessed through a common methodology, be collected by a group representative of the national population, and conform to international standards. In 2006, the United States Institute of Medicine issued guidance on the selection of valid indicators (Institute of Medicine, 2006). It identified good performance indicators as being scientifically sound, feasible, important, aligned across datasets, and comprehensive. In 2008, the Association of Public Health Observatories and the UK's NHS Institute for Innovation and Improvement published "The Good Indicators Guide" (Pencheon, 2007). In this guide, the authors set forth a set of criteria for good performance indicators, including: importance and relevance, validity, possibility, meaning, and implications. Several of these criteria map well to many of the challenges of healthcare evaluation described by Ozcan (2005). For example, the Guide's validity criteria ensure that the indicators account for patient participation and medical heterogeneity. Armed with well-defined guidelines, the process of selecting indicators became more systematic and, consequently, robust (Boulkedid et al., 2011).

As this area of research matured and performance indicators were tested and implemented in real-world settings, more specific, detailed, and practical guidance was issued for choosing the best performance indicators and using them appropriately. Wollersheim et al. (2007) published a review of best practices for the selection and application of performance indicators. The review included practical advice, such as a seven-step plan for implementing a performance indiactor system, specific databases to search for finding previously implemented performance indicators, a recommended structure for performance indicator selection committees, and vivid examples of the development and use of performance indicators in three clinical contexts. The authors emphasized that, when developing and selecting performance indicators, selection commitee should first survey measures of quality in a broad set of clinical domains, match the indicators to the evidence base, and then prioritise the indicators through a consensus model with experts. In 2010, The Kings Fund issued a report on practical issues of the selection and use of performance indicators (Raleigh and Foot, 2010). In this report, the authors advised that only the best performance indicators be adopted into practice. They also advised using the Donabedian Model (Donabedian, 1980) to choose indicators appropriately balanced among the structure, process, and outcome domains. Using this framework, they raised awareness of potential pitfalls for indicators in each domain; for example, they describe that process indicators, while more responsive and less prone to several forms of bias than outcome indicators, were susceptible to gaming and may require new forms of data collection. The Kings Fund report further explained how indicators should be analysed. In particular, the authors strongly recommend that analysis of healthcare quality measures be adjusted for case-mix and determinants of health outside of the healthcare services industry.

Opinions splintered around the concept of thresholds. Traditionally, performance indicators were designed such that a given practice's performance on a specific metric was measured against an *a priori* fixed threshold(s). Performance indicators with fixed threshold – also known as absolute targets or criterion referenced targets – have several advantages (Doran et al., 2014). For example, physicians can develop specific strategies to reach a set target. Further, physicians can be assured that their efforts, if successful, will not go unrewarded. However, absolute targets also have some severe disadvantages. Healthcare payers cannot know their costs upfront; if more or less physicians achieve the target than previously expected, it could lead to a budget surplus or deficit. Further, quality monitors may not have complete information on the full capabilities of physicians and, consequently, may set the thresholds too low or too high. To ameliorate these issues, some began advocating for relative quality measures, in which a healthcare provider received a reward depending upon their place in the distribution of all providers for a particular metric (Chien and Rosenthal, 2013). Notably, relative thresholds were adopted in the United States Medicare program through the Physician Value Based Payment Modifier provision of the Affordable Care Act. Some have advocated for a middle ground; in 2009, NICE recommended fixing performance targets to different points of the previous year's distribution of performance (Doran et al., 2014). In this way, practices would know *a priori* the targets they need to achieve, but thresholds would have a reasonable, empirical basis.

Despite the depth and rich history of the literature on performance indicators, there is no universally accepted method or international standard for developing performance indicators (Shekelle, 2013; Stelfox and Straus, 2013). Some approaches vary only semantically, while others have more substantial differences. For example, Stelfox and Straus (2013) take a population-health approach by recomending that quality indicators be restricted to clinical problems with "large burden of illness to justify quality measurement and improvement efforts". Shekelle (2013), on the other hand, takes a more individual-centric approach, noting that while quality of care for rare diseases may not meaningfully affect population health, it is only equitable that those unfortunate enough to have rare diseases receive the same protection from performance indicators. Instead, individual institutions need to make their own "judgement calls" to determine which best practices from this literature work best in their own clinical, organisational, and political context.

1.3 The Quality and Outcomes Framework

1.3.1 Selecting Performance Indicators for UK General Practitioners

In this vein, the UK did not adopt a formal performance indicator selection protocol prescribed by the literature in its Quality and Outcomes Framework pay-for-performance scheme, opting instead to form their own. However, common themes of the discourse related to selection and analysis of healthcare performance indicators laid the foundation for the initial implementation and subsequent restructures of the Quality and Outcomes Framework scheme, and references to this literature (both explicit and implicit) can be observed in QOF policy documents and in the QOF structure.

The first set of indicators were negotiated by representatives from the British Medical Association and the NHS over an 18-month period (Doran et al., 2014). While recollections of these initial negotiations are sparse in the literature, there were some set ground rules for developing the indicators, and these ground rules appear to be linked with the preceding literature. For example, GPs were given the ability to exclude certain patients from their indicators based upon refusal by the patient or clinical inappropriateness (Roland and Guthrie, 2016), mitigating concerns that patient participation and heterogeneity would interfere with quality measures. Results from patient experience surveys were included as key measures of quality, emphasizing that the QOF would take a comprehensive approach to quality. The selected indicators were designed with a fixed threshold(s), as was most common in the literature.

By 2005, the performance indicators were selected more systematically. That year (and again in 2007), in line with the literature's emphasis on comprehensive performance indicators, topics for development were garnered from a wide swath of stakeholders, including physicians, patients, non-profit organisations, governmental agencies, and private industry (Lester and Campbell, 2010). These topics were then matched to the evidence base and prioritised by an expert panel including representatives from the British Medical Association's General Practitioners' Committee and the Department of Health. In 2007, the indicator selection protocol also included meetings with groups that had made suggestions that were ultimately prioritised and the use of a modified Delphi procedure, a formal consensus system based on both scientific literature and expert opinion.

Since 2009, the UK Government's National Institute for Health and Clinical Excellence (NICE) has been formally tasked with managing the process by which performance indicators for the QOF scheme are developed (Sutcliffe et al., 2012). While the final indicators are still selected through a negotiation between the NHS the British Medical Association's General Practitioners' Committee, the NICE process has effectively set the agenda for which performance indicators should be added, retained, or removed each year by developing a menu of "appropriate" or "suitable" indicators (NICE, 2019).

Initially, NICE quality indicators were developed through a structured process led by a NICE-supported but formally independent Quality Standards Advisory Committee, which was made up of 21 standing members (including physicians, administrators, public health officials, and laypeople) and five specialists (Bennett et al., 2014). Today, the process is led by a similar group called the NICE Indicator Advisory Committee (National Institute for Health and Care Excellence, 2017). In the first stage of the process, referred to as "Topic Overview", NICE staff engaged with topic-specific stakeholders for two weeks and prepared briefings on the areas discussed, using the Donabedian conceptual model to guide their exploration (Donabedian, 1980). Then, in the "Prioritising Quality Areas" stage, the Advisory Committee received the NICE briefings and met to agree on topic-specific priorities. Priority was determined according to (A) evidence that quality of care varies across GPs, (B) feasibility of improving quality in the area, and (C) possibility of constructing a valid measure around that topic. Specifically, the committee is advised to rate proposed indicators against the following criteria:

- The proposed indicator should focus either on a health outcome or on a process that is closely linked to improved outcomes.

- The proposed indicator should focus on a national priority, particularly one that is specifically identified by either NHS England, Public Health England, or the devlolved administrations.

- The proposed indicator should be feasibly measurable, either through existing or possible data collection systems.

- The proposed indicator should relate to an area where there is currently variation in practice and evidence that adoption of best practices can improve outcomes.

- The proposed indicator should reflect upon the actors it intends to measure (e.g., general practitioners).

Although not formally recognized in NICE documentation, scholars and regulators have commented that additional criteria apply to the selection of QOF indicators (Campbell and Lester, 2010; Gillam and Steel, 2013). They suggest that a clinical area's "QOFability", or suitability for inclusion in the QOF program, relate to (A) its prevalence in the population, (B) its effect on morbidity and mortality, and (C) the ease with which it is measured. Further, specific indicators are deemed more "QOFable" if they are easily accessed from QMAS, based on Government guidelines, well defined, universally achievable, unlikely to be "gamed", and directly related to the GP's actions. For example, indicators involving medical procedures and drug interventions may be prioritised in the QOF over community health interventions simply because the former are more straightforward than the latter (Gillam and Steel, 2013)

With priorities set by the advisory committe, NICE and the NHS begin the process of drafting the indicator. The two agencies (while accepting public consultation) to define the indicator, define a methodology that could be used to collect data on the indicator, perform feasibility and/or pilot testing on the indicator, assess cost-effectiveness of adopting the indicator, and finally conduct a review of possible thresholds for the indicator (Campbell et al., 2011). In cases where a relevant indicator already exists, it is reviewed in this stage, as well. Through all of these steps, NICE arrives at an indicator statement, which is then subject to a 4-week public review by stakeholders and members of the public. The draft statement and stakeholder comments is then presented to the Advisory Committee, which made necessary adjustments. Next, NICE staffers checked the validity of the statement and its consistency with other indicators. The final product was then passed along to the NICE Guidance Executive, a group of NICE directors, who gave final approval for the indicator statement. After a performance indicator was approved by NICE, it invariably became the subject of future academic and policy research, which can then be used to change or eliminate the indicator.

Once all NICE indicators are approved through this process, they are published on NICE's website (along with any related information collected during the selection process) as suitable for implementation. Importantly, NICE performance indicators are not automatically implemented in the QOF scheme but, instead, only enter the QOF scheme through contractual negotiation between the NHS and the British Medical Association's General Practitioners Committee. That is, NICE indicators are implemented (and corresponding point values set) when a new contract is agreed upon by the NHS and GPs (NHS England, 2019). Consequently, through the contract negotiation process, QOF performance indicators are subject to an additional political feasibility constraint beyond NICE's recommended indicators.

1.3.2 The Purpose of the QOF: Pay Raise or Incentive to Improve?

An important quality of any performance metric is that it be aligned with overarching goals. Unfortunately, while the QOF's focus on quality and outcomes is clear, its overall purpose has not always been well defined. Consequently, it is important to note that the evaluation of any indicator may depend partially on the evaluator's own view of the purpose of the QOF scheme.

At the time the QOF was introduced, it was widely regarded as a bargaining chip for increased GP pay. The NHS had already agreed that an increase in pay was necessary to sustain the GP workforce, and participation in a pay-for-performance scheme was an at least politically necessary *quid pro quo* (Roland and Guthrie, 2016). It remained ambiguous whether the QOF scheme was simply a means to justify a pay raise – intended to reward GPs for continuing the work they were already doing – or whether the QOF was intended to become a mechanism to incentivize GPs to adopt higher-quality practices in their clinics. From the initial indicators selected, it appeared that the QOF was simply a mechanism to systematically observe GP quality while offering them higher pay. According to Roland (2007), most of the original indicators were based on existing guidance, such as the National Service Frameworks, and within the first year of implementation, GPs earned 91.3% of total QOF points (Dixon et al., 2011).

Subsequently, however, some indicators that were not already common practice were adopted in order to incentivize changes in practice (Roland and Guthrie, 2016). By 2006, the NHS stated that "[t]he QOF is not about performance management of general practice but about resourcing and then rewarding good practice." (National Health Service, 2006*a*). Since the 2009 QOF scheme reforms, NICE has become more clear that the QOF indicators are intended to incentivise quality improvement rather than reward the status quo. For example, NICE asserts that its quality indicators are developed with the "aim to shape measureable quality improvements by identifying the key areas for improvement within a particular area of health, public health or social care" (Bennett et al., 2014). Further, one of the criteria NICE uses to select performance indicators relates to the feasibility of improvement (National Institute for Health and Care Excellence, 2017).

Whether this is true, in practice, however, is debatable. As of 2013, most indicator's upper thresholds were held below the average practice's achievement rate (Doran et al., 2014), and the average practice participating in the QOF scheme achieves over 90% of available points.

1.3.3 The QOF Scheme in Practice

When the QOF scheme was introduced as the largest pay-for-performance scheme in the world in 2004, it linked 146 indicators across four domains (clinical, organisational, patient experience, and additional services) with approximately 20-25% of overall practice income (Guthrie, McLean and Sutton, 2006). By participating in the (voluntary) program and meeting thresholds within these indicators, practices had the opportunity to earn up to 1050 total points. At the end of the year, point totals would be assessed and practices would be commensurately renumerated, with marginal adjustments made for workload and patient health. For reference, the average practice could expect to recieve approximately 77 GBP for each point it earned (see Table 4).

The Clinical Domain commanded the greatest weight in the QOF with 76 indicators across 11 areas holding an allocation of 550 points (52.4% of total points). The organisational domain was the next most significant, with 56 indicators across five areas holding an allocation of 184 points. The patient experience domain had just four indicators across two areas but with an allocation of up to 100 points. The Additional Services domain, which included ten indicators over four areas, held the least weight in the scheme, with an allocation of only 36 points. In addition, there were some additional opportunities for practices to earn points, often called "depth of quality measures". A holistic care measure, which assessed achievement over the entire clinical domain, was worth up to 100 points. A quality practice measure, which assessed overall achievement in the non-clinical domains, was worth up to 30 points. Finally, an access measure, which assessed patient access to clinical care, was worth 50 points. The areas assigned to each of the four domains are shown in Table 1.

Most QOF indicators had fixed lower and upper thresholds related to the percentage of eligible patients for whom an indicator was achieved (Doran et al., 2014). If a practice's performance was below the lower threshold, it would receive no QOF points and no renu-

Table 1: Domains and Areas of 2004-2005 QOF Scheme

Clinical Domain	**Organisational Domain**
Total Avaialble Points: 550 (52.4%)	Total Avaialble Points: 184 (17.5%)
Coronary heart diseases	Records and information
Left ventricular dysfunction	Patient communication
Stroke and transient ischaemic attack	Education and training
Hyptertension	Medicines management
Diabetes mellitus	Clinical and practice management
Chronic obstructive pulmonary disease	
Epilepsy	
Hypothyroidism	
Cancer	
Mental health	
Asthma	
Patient Experience Domain	**Additional Services Domain**
Total Avaialble Points: 100 (9.5%)	Total Avaialble Points: 36 (3.4%)
Patient survey	Cervical screening
Consultation length	Child health surveillance
	Maternity services
	Contraceptive services

NOTE: The data for this table was reported by Lester and Campbell (2010). Total Available Points do not add to 1050 (100%) because 180 points are designated to "depth of quality measures" outside the four domains.

meration. If it met or exceeded the upper threshold, it would receive the maximum QOF points and payment. If the practice's performance fell between the lower and upper thresholds, it would receive points and payment between the minimum and maximum amount. All QOF lower thresholds were set at 25% of eligible patients, while upper thresholds ranged from 50% to 90% depending upon the perceived difficulty of achieving the indicator. Other indicators took on binary values, for which practices received full points when achieved and 0 points if not achieved. Most indicators had a timeframe of 15 months. Table 2 provides several examples of indicators from the 2004-2005 QOF scheme, including their point value, lower threshold, and upper threshold.

While points achieved were directly linked with payment, the QOF scheme had a few nuanced features that modified the relationship between points and payment through practice-level clinical condition prevalence Dixon et al. (2011). For each clinical outcome, payment was calculated using the square root of prevalence among the practice's registered patients. For this reason, practices with high disease prevalence received lower payment per patient treated to the indicator's standard than practices with lower disease prevalence. Further, those practices with very low disease prevalence (under 5% of the maximum) were rewarded as if they had a higher prevalence even though they had fewer patients to care for. These adjustments disproportionately hurt smaller practices from areas with lower average health relative to larger practices with higher average health.

The overall structure of the QOF scheme (i.e., practice-level pay-for-performance based upon predetermined threshold values) has remained mostly constant since its inception, but changes have been made to the domains, areas, indicators, and thresholds. A review of changes to the QOF for England is summarised in Table 3.

Table 3: Timeline of Changes to the QOF Scheme (England)

Contract Year	**Summary of Changes**
2004/2005	First year of QOF
2005/2006	Average payment per QOF point increases from £77.50 to £124.60

Contract Year	**Summary of Changes**
2006/2007	Lower thresholds raised from 25% to 40%. Some upper thresholds changed. 166 QOF points redistributed. 50-point access measure removed, reducing total QOF points to 1000.
2007/2008	
2008/2009	58 QOF points reallocated to an indicator that patients have access to a healthcare professional within 48 hours (PE7).
2009/2010	72 QOF points reallocated. Clinical area for cardiovascular disease primary prevention added, and new indicators were introduced for heart failure, chronic kidney disease, depression, diabetes areas. Two high-value patient experience indicators were removed. Prevalence adjustment (also known as "square rooting") removed.
2010/2011	No changes due to Swine Flu.
2011/2012	Indicator that patients could access a healthcare professional within 48 hours (PE7) was removed. Upper thresholds for two indicators were increased by one percentage point.
2012/2013	New clinical areas for peripheral arterial disease and osteoperosis. Seven indicators were retired, and 12 new indicators were introduced. Lower thresholds were increased by 5-10 percentage points for all indicators, and upper thresholds were increased by 4-10 percentage points for 13 indicators.
2013/2014	Devolved governments begin negotiating their own QOF schemes. Upper thresholds were set at the 75^{th} percentile of achievement and lower thresholds were set at 40 percentage points below the 75^{th} percentile for 20 indicators. Base payment per point was increased from £133.76 to £156.92. The organisational domain was removed, and new domains were added for "Public Health" and "Quality and Productivity" (the Additional Services domain was renamed "Public Health – Additional Services"). The timeframe for most indicators was reduced from 15 months to 12 months. The total number of available QOF points was reduced from 1000 to 900.
2014/2015	Several unpopular indicators were removed. The "Quality and Productivity" and "Patient Experience" domains were removed, as well as the hypothyroidism, child health surveillance, and maternity areas. Age restrictions for learning disability indicators were removed, while those for blood pressure were raised from 40+ to 45+. The total number of available QOF points was reduced from 900 to 559.
2015/2016	Upper thresholds were set at the 75^{th} percentile of achievement and lower thresholds were set at 40 percentage points below the 75^{th} percentile for remaining indicators. The average payment per QOF point was increased to £160.15. Minor changes were made to the atrial fibrillation, coronary heart disease, dementia, chronic kidney disease, and obesity areas.
2017/2018	No changes
2018/2019	No changes

Contract Year	Summary of Changes

NOTE: These changes are adapted from Institute of Healthcare Management (2018) with additions from Doran et al. (2014), Stokes (2014), and National Health Service (2014). For years 2004/2005 to 2012/2013, the QOF point universally applied to England, Wales, Scotland, and Northern Ireland. Those changes from 2013/2014 on apply only to England.

Table 2: Sample Performance Indicators and Point Value, QOF 2004-2005

Indicator	**Point Value**	**Lower Threshold**	**Upper Threshold**
CHD 7. The percentage of patients with coronary heart disease whose notes have a record of total cholesterol in the previous 15 months	7	25%	90%
LVD 1. The practice can produce a register of patients with CHD and left ventricular dysfunction	4	NA	NA
BP 2.The percentage of patients with hypertension whose notes record smoking status at least once	10	25%	90%
MH 2. The percentage of patients with severe long-term mental health problems with a review recorded in the preceding 15 months. This review includes a check on the accuracy of prescribed medication, a review of physical health and a review of coordination arrangements with secondary care	23	25%	90%
EPILEPSY 4. The percentage of patients age 16 and over on drug treatment for epilepsy who have been convulsion-free for last 12 months recorded in last 15 months	6	25%	90%
INFORMATION 1. The practice has a system to allow patients to contact the out-of-hours service by making no more than two telephone calls	0.5	NA	NA
EDUCATION 4. The practice has a system to allow patients to contact the out-of-hours service by making no more than two telephone calls	3	NA	NA

NOTE: These examples are reported from National Health Service (2005). Those indicators with an NA upper and lower threshold are binary indicators, where the practice receives full points if the indicator is satisfied and 0 if the indicator is not satisfied.

Table 4: Average Payment Per QOF Point (England)

Contract Year	**Average Payment Per Point**	**Source**
2004/2005	£77.50	Institute of Healthcare Management (2018)
2005/2006	£124.60	Gillam and Siriwardena (2011)
2006/2007		
2007/2008		
2008/2009		
2009/2010		
2010/2011	£127.29	Institute of Healthcare Management (2018)
2011/2012	£130.51	Institute of Healthcare Management (2018)
2012/2013	£133.76	Doran et al. (2014)
2013/2014	£156.92	Doran et al. (2014)
2014/2015		
2015/2016		
2017/2018		
2018/2019	£179.26	Institute of Healthcare Management (2018)

NOTE: For years 2004/2005 to 2012/2013, the QOF scheme universally applied to England, Wales, Scotland, and Northern Ireland. Those values from 2013/2014 on apply only to England. Payment per point is subject to adjustment for practice demographics; before 2009/2010, QOF payments per point was subject to "square rooting".

Perhaps the most dramatic revision of the QOF program occurred with the 2006 General Medical Services contract at the behest by the newly formed Expert Panel. Most lower thresholds for clinical indicators were raised from 25% to 40%, and a few upper thresholds were also adjusted (12 of 50 were raised by 5-20 percentage points and 1 was lowered by 10 percentage points) (Doran et al., 2014). Further, seven new clinical areas were added: dementia, depression, chronic kidney disease, atrial fibrillation, palliative care, obesity and learning disabilities (Applebee, 2006). The left ventricular dysfunction area was replaced by heart failure. The total available QOF points was reduced from 1050 to 1000, as the 50 point access measure was moved to the Directed Enhanced Services scheme, an optional program for practices to meet the needs of their local communities (National Health Service, 2006*b*). Indicators within existing clinical areas were amended, and point values were moved around, both within the existing indicators and from existing areas to indicators in new areas. The

2006 revision of the QOF scheme was seen as an increased burden on GPs. Even in cases where indicators had been removed in the revision, the contract negotiators had stated that those indicators were now part of standard, "good medical practice" (Applebee, 2006).

A minor revision of the QOF scheme occurred in 2009, when NICE took leadership over the indicator selection process. A new clinical area for cardiovascular disease primary prevention was introduced, and new indicators were added to the heart failure, chronic kidney disease, depression, and diabetes clinical areas (National Health Service, 2010). Additional changes were made to the patient experience and additional services domains. Importantly, the prevalence adjustment mechanisms were removed, which ensured the QOF would fairly reward practices in lower average health areas (Dixon et al., 2011).

In 2010, NICE recommended benchmarking thresholds against the distribution of the past year's performance, but this suggestion was rejected during contract negotiation (Doran et al., 2014). Instead, both parties agreed to minor increases in the upper thresholds in 2011/2012 and slightly more substantial increases to lower and upper thresholds in 2012/2013. Minor changes to the indicators also occurred, including the addition of two new clinical areas for PAD and osteoperosis.

The next notable change for the QOF scheme occurred in 2013, when it was decided that devolved governments would negotiate their own QOF contracts. From the 2013/2014 contract forward, the QOF scheme for England, Wales, Scotland, and Northern Ireland were permitted to vary (Institute of Healthcare Management, 2018). England made a number of significant changes, including the addition of "Public Health" and "Quality and Productivity" domains and the removal fo the "Organisational" domain. A reduction of the total number of available QOF points occurred in line with an increase in the average payment per QOF point. England also adopted a system in which upper thresholds would be set at 75^{th} percentile achievement from the previous year (Doran et al., 2014). To give GPs time to adapt to changes in thresholds, upper thresholds were changed for only 20 indicators in 2013/2014, with the remainder indicators' upper thresholds changed in 2015/2016 (postponed from 2014/2015).

Another set of changes occurred in 2014/2015. The “Quality and Productivity” and “Patient Experience” domains were removed, leaving just the “Clinical”, “Public Health”, and “Public Health – Additional Services” domains. The total number of available QOF points were reduced from 900 to 559.

Importantly, these changes to the QOF scheme corresponded with variation in the balance of weight between domains (Table 5) and the average achievement rate of practices (Table 6).

Table 5: Distribution of Points Available across Domains, 2006-2012

Contract Year	Additional Services	Clinical	Organisational	Patient Experience	Total
2006/2007	3.6%	65.5%	18.1%	10.8%	100.0%
2007/2008	3.6%	65.5%	18.1%	10.8%	100.0%
2008/2009	3.6%	65.0%	16.8%	14.6%	100.0%
2009/2010	4.4%	69.7%	16.8%	9.2%	100.0%
2010/2011	4.4%	69.7%	16.8%	9.2%	100.0%
2011/2012	4.4%	66.1%	26.2%	3.3%	100.0%

Table 6: Average Achievement across Domains, 2006-2012

Contract Year	Additional Services	Clinical	Organisational	Patient Experience	Total
2006/2007	96.5%	96.3%	92.5%	95.9%	95.5%
2007/2008	97.0%	97.5%	94.5%	97.2%	96.8%
2008/2009	97.3%	97.8%	95.8%	84.2%	95.4%
2009/2010	95.3%	95.9%	96.3%	71.5%	93.7%
2010/2011	97.1%	96.8%	97.4%	72.6%	94.7%
2011/2012	97.0%	97.0%	96.4%	99.0%	96.9%

As of 2018/2019, the QOF remains integral part of general practice in the UK. While the program is still voluntary, over 95% of GP practices (N=6873) participated with an average achievement score of 539.2 points (of 559 possible) and 13.0% achieving the maximum score (National Health Service, 2019).

1.4 Social Influence in General Practice Quality

1.4.1 Introduction to Social Influence

Social influence may be an important and understudied determinant of quality in healthcare. The notion that physicians may influence the quality of the healthcare others deliver through social influence[2] is borne from two existing literatures. The first field demonstrated that, while successful quality improvement interventions are rare (Conry et al., 2012), physicians respond well to information delivered by peers and peer reviews of their practice. The second field documented a specific pattern of medical practice variation in which variance is lower within groups than between groups. Peer influence was proposed as one possible mechanism for this pattern. To date, only a few studies have attempted to test the role of social infleunce, primarily using data on physicians who practice in multiple locations. If social influence has a significant impact on clinical decision making in regards to healthcare quality, it could become a useful policy lever for quality improvement.

1.4.2 Peer Influence in Quality Improvement

Peer influence, encompassing clinical peer review, peer feedback, and peer teaching, is an established vehicle for quality improvement (Sargeant et al., 2015). In fact, the Joint Commission on Accreditation of Healthcare Organizations required physician peer review as part of its accreditation process as far back as 1952 (Goldberg, 1984). Since at least the 1970s, robust studies have documented the efficacy of peer-based interventions in improving quality in several domains of healthcare delivered by general practitioners and other physicians (Mittman, Tonesk and Jacobson, 1992; Jamtvedt et al., 2003). Today, "peer review and clinical audit" is two of the key strategies for quality improvement endorsed by the World

[2] The research literature has frequently used the terms "peer influence" and "social influence" interchangably. For purposes of clarity in this dissertation, I distinguish between these terms. "Peer influence" will refer to the influence that all similar physicians hold over an individual physician, whereas "social influence" refers to how direct interactions with other physicians may affect a physician's behavior. For example, peer review by a government body of physicians or a formal, instructional visit led by another physician would fall under the category of "peer influence". Influence among physicians who are friends or work in the same practice – those who interact often and on an on-going basis – would be considered "social influence".

Health Organization (World Health Organization, OECD and International Bank for Reconstruction and Development/The World Bank, 2018).

Several studies and reviews have demonstrated that peer-based interventions are successful method of improving quality in general practice (Wensing, Weijden and Grol, 1998). For example, peer-based interventions have successfully improved preventive medicine in general practice (Hulscher et al., 1999). Using a randomised controlled trial design, Inui (1976) found that a tutorial session improved dietary counseling and hypertension monitoring practice by 44% relative to a control group with no intervention. Similarly, Cockburn et al. (1992) reported that, in a randomised controlled trial, an instructional visit by a facilitator or a courier improved GPs adherence to smoking counselling guidelines by 15% and 5% relative to sending the information by mail. Relatedly, several studies have shown that peer-to-peer interventions can be successful in reducing overprescribing rates (Arnold and Straus, 2006; Avorn and Soumerai, 1983). For example, a multi-institutional, quasi-experimental study of Mexican physicians found that an intervention involving a managerial peer review committee improved prescribing practices among up to 40% of physicians (Pérez-Cuevas et al., 1996). In a parellel randomised controlled trial using peer feedback groups to address overprescribing by Dutch GPs in the treatment of asthma and urinary tract infection found that peer groups significantly improved clinical practice (Veninga et al., 2000). Peer-based interventions have also shown efficacy in reducing unnecessary referrals to secondary care. Six weekly “cluster group” meetings were able to successfully reduce NHS GP specialist referral from 5.5 patients per 1000 to 4.3 patients per 1000 (Evans, Aiking and Edwards, 2011).

Peer influence is also a well-tested and evidence-based vehicle for quality improvement for several medical fields outside of general practice. A study of the Mayo Clinic Health System found that the implementation of clinical peer review reduced the all-cause mortality rate in a critical access hospital by 50% (Deyo-Svendsen et al., 2016). Two studies found that programs involving peer-to-peer site visits reduced mortality associated with coronary artery bypass grafting (Holman et al., 2001; O’Connor et al., 1996). A common and well-studied form of peer-based interventions to improve healthcare quality, particularly within specialty care, are Communities of Practice (also known as clinical communities) (Aveling et al., 2012; Endsley, Kirkegaard and Linares, 2005; Iedema, Meyerkort and White, 2005). Generally self-organised groups of physicians in the same specialty, Communities of Practice share relevant clinical information and form peer review practices. Communities of Practice may involve interactive training sessions, story-sharing, and promotion of internal audits. These Communities have demonstrated efficacy for improving clinical practice. For example, a Michigan-based Community of Practice intervention reduced the number of infections per catheter-days in Intensive Care Unit settings by more than 65% (Pronovost et al., 2016).

Collectively, the empirical evidence in favor of peer-based quality improvement interventions has made these among the most popular in medicine. Scholars have suggested several theoretical reasons for the success of these interventions. Some suggest that the vividness and practical applicability of story-telling may play a role; physicians may communicate with one another in a narrative form that other physicians can directly apply to their practices (Jordan et al., 2009). Others hypothesize that peer-based interventions encourage a level of interactivity that facilitate learning (Johnson, Manyika and Yee, 2005) or that the trust in peer-relationships facilitate adoption (Wenger, 1999). Recently, some have suggested that peer-based interventions could lead to consistent relationships that impose social influence (Meltzer et al., 2010).

1.4.3 Within-Group and Between-Group Medical Practice Variation

Largely separate from the literature on healthcare quality improvement interventions, studies stretching over several decades have found consistent patterns of Medical Practice Variation in the UK. In the 1930s, Dr. J. Alison Glover reported in *Proceedings of the Royal Society of Medicine* that rates of tonsillectomies among children varied significantly among different geographic regions (Glover, 1938). Subsequent studies of small area variation in medical practice extended this finding to several areas of medical practice and surgery (Wennberg and Gittelsohn, 1973, 1982; Paul-Shaheen, Clark and Williams, 1987; Jaffer et al., 2010; O'Connor et al., 1999).

More recently, researchers have used microlevel diagnosis and treatment data and multilevel modelling techniques to document a more essential feature of Medical Practice Variation: variance in medical practice tends to be greater between rather than within medical groups. Westert, G.P. (1992), for example, found that duration of hospital stays was more similar within hospital wards than across wards. This pattern in variation in duration of hospital stay has now been observed in several settings (van de Vijsel, Heijink and Schipper, 2015). Using data from the Dutch National Survey of General Practice and techniques from social network analysis, de Jong, Groenewegen and Westert (2003) found that GPs tend to be more similar within partnerships than between partnerships. For example, they found that self-reported practice behavior (medical techniques used, handling work style elements), time spent on clinical activities, and practice choice in response to a given complaint were more similar within practices than between practices.

Several theories were posed to explain this pattern of within versus between variation. Some suggested that selection of doctors to practices may play a role. Similar people tend to attract (Fehr, 1996), and physicians may choose to work with and recruit other physicians who are similar to them. Others suggested that (observed or unobserved) differences in circumstance may vary practice between but not within practice. Physicians within a given practice presumably have the same incentive structure and the same access to information and resources (de Jong, Groenewegen and Westert, 2003), but this need not be the case for physicians in different practices. If, for example, physicians in practice A is judged on a different metric from those in practice B, physicians may respond accordingly (Krasnik et al., 1990) and exhibit medical practices that vary markedly between practices. However, all physicians within practice A or within practice B would be aligned.

Social influence was also hypothesized as an explanation of this pattern. Sociologists have observed that people desire social acceptance, avoid deviation from and adapt towards social norms (Eddy, 1984; Dambrun, Guimond and Duarte, 2002), and evaluate their own performance in relation to role equivalents (Fehr, 1996; Erickson, 1988). In this vein, "people both shape and are shaped by social networks", even when adapting to others conflicts with rational choice (Pescosolido, 1992). Further, social cues may be a vehicle for physicians to avoid risk in the context of uncertainty (Bandura, 1986). Indeed, physicians report that "informal consultations" with other physicians are common (Keating, Zaslavsky and Ayanian, 1998) and even that they may affect practice and quality of care (Eisenberg, 1985; Keating et al., 2007; Geneau et al., 2008). Consequently, physicians may "follow the pack" with the understanding that deviations from the practices of their proximate peers could reduce social approval or increase risk (Westert and Groenewegen, 1999).

1.4.4 Empirical Evidence of Social Influence Effect

Only a small number of studies have adequately tested the hypothesis that social influence among physicians may affect their practice. These studies typically evaluate how physicians who practice in multiple locations vary their practice based upon location. In this way, the researchers exclude the possibility that how physicians select into practices impacts the within-group variance; because the physician is the same person in both settings, selection can be ruled out. Further, this approach mitigates (but does not eliminate) the possibility that differences in circumstance explain the pattern of within versus between group variation. It is true that a physician could work in different practice settings which, themselves, have many different circumstances, such as equipment availability, support staff, or incentive schemes. However, given that the physician is the same in both practice settings, the knowledge and intrinsic resources of the physician should be the same. Consequently, while acknowledging the limitations, a comparison of how physicians who practice in multiple locations vary their practice based upon location is a reasonable estimate of the impact of social influence on medical practice.

The earliest example comes from Griffiths, Waters and Acheson (1979), who studied physicians who performed elective repair of inguinal hernia in multiple hospital settings within the NHS. They found that physicians who performed the procedure at more than one hospital significantly varied the patient's length of postoperative stay among the hospitals, while physicians who performed the procedure at just one hospital were relatively consistent. Westert, Nieboer and Groenewegen (1993) applied modern statistical approaches in a similar setting in the Netherlands and found similar results; physicians who performed similar procedures across similar patients at more than one hospital varied their patients' length-of-stay to match each hospital. Subsequently, Jong et al. (2006) found a similar pattern for length-of-stay among multi-hospital physicians practicing in the United States. To date, no study involving doctors in multiple locations has evaluated the effect of social influence on general practitioners.

Apart from these, only a few studies have even hinted at an effect of social influence among physicians. Chung et al. (2003) found that new physicians who spent more time in the same residency (i.e., training) rotation shared more similar practice styles. While this points to social influence, it is possible that rotations were subject to differential lessons that were reinforced over time. Goodwin et al. (2013) studied the effect of hospitals on length of stay and discharge destination among patients in Texas. In a two-level model involving admissions and hospitalists, there was significant variation in practice among hospitalists. However, this variation was greatly attenuated with the inclusion of a third level: hospitals. In this way, the authors find that the hospitalist-level variation is largely explained by the hospitals within which the hospitalists' practice. Again, it is possible this is indicative of a social influence effect; however, the selection or circumstance hypotheses offer alternative explanations.

1.5 Implications of A Mover Effect

1.5.1 Leveraging the Social Influence Effect

If social influence is determined to be a meaningful and significant driver of medical practice variation, it could become a powerful policy lever for improving healthcare quality in general practice. Selection of certain GPs into practices is challenging to overcome without an administrative take-over of general practitioners offices, and while some variation in circumstance (e.g., incentive schemes) can be managed, others (e.g., availability of equipment or information) can be substantially more difficult to mitigate. Several actionable steps can be taken by both the government and individual GP practices in order to leverage social influence.

Quality improvement programs should focus on not just giving GPs information on best practices. Instead, these programs should focus on shifting social norms towards those best practices.

The Government may consider instituting rotational programs or sabbaticals that expose people in worse performing practices to those in better performing practices. For example, the Government may incentivise GPs who recently left top-performing areas to spend a few months at a time in low-performing practices, with the hope that the newcomer will use her social influence to shift quality of practice upwards.

Similarly, the Government may consider incentivising permanent moves to lower performing practices. Previous research suggests that physicians would be willing to move locations for an incentive (Scott et al., 2013). If a GP mover is able to affect the quality of her incoming practice, the benefits to that practice may outweigh the costs of an incentive program.

GPs may wish to consider the social influence of candidates when hiring a new GP. These practices may wish to hire GPs from higher-performing practices in the hopes that the newcomers' practices will shift the social norm towards higher quality of care.

Conversely, if a GP practice is hiring an individual from a worse performing practice, the existing partners may want to emphasize the need for the newcomer to quickly adapt to the practice's standards.

1.5.2 Credibility of Quality Measures

While quality measures like those in the QOF scheme are commonplace in healthcare, they are not universally accepted. Some feel strongly that the quality measures are invalid or unreliable. For example, the "Altruism" model poses that practices perform in accordance with their available resources and information. In this case, rewarding high-performing practices is counterproductive; QOF payments are undeserved by GPs and fail to help low-resource, low-performing practices improve.

If there is an observable "mover" effect, however, this strongly suggests that the makeup of physicians within a given practice determine its performance. If the addition of one GP from a better or worse practice into the target practice can affect the target practice's quality scores, then the quality scores are a function of the GPs.

It is possible that a "mover" effect exists within the context of the "Altruism" model – for example, if a doctor moves in with better information, that would elevate the information of the practice. However, in this case, a mover from a worse practice should not meaningfully affect quality; the new person would benefit from the shared information in the practice, which would stay constant. Therefore, if there is a *negative* effect from GPs entering from worse practices, this is unaccounted for in the "Altruism" model. Consequently, the presense of a "mover" effect enhances the credibility of quality measures.

2 Factors Influencing GP Movement

There are several factors that could potentilly influence GP movement from one practice to another, and it is important to consider these factors as mechanisms that could be driving the "mover effect". The factors influcing GP movement be divided into four main categories:

Table 7: Factors Driving GP Movement

	Any Move	**Destination**
GP	What Prompts a GP to Want to Move	What Determines Where the GP Moves
Practice	What Prompts a Practice to Hire a New GP	What Does the New Practice Look for in a New GP

NOTE: This table was developed by the author.

factors that prompt a GP to move, factors the GP considers when choosing where to move, factors that prompt a practice to hire a new GP, and factors that the practice considers when choosing a new GP (see Table 7). There are some studies that consider which factors are most important in setting GP's preferences for practices (i.e., the top-right box of Table 7), but the empirical and theoretical literuatre on the other categories is limited.

2.1 Physician Preferences for Practices

One of the first studies to consider the factors that drive physicians' movement within a developed country was conducted by Beardow, Cheung and Styles (1993). The authors surveyed GP trainees in the UK on which factors influenced their decision of practice within the North West Thames Regional Health Authority, finding that the most important factors were (A) a good relationship with existing staff and GPs at the practice, (B) whether the practice had a practice nurse or practice manager, (C) whether the practice had attached health authority staff, (D) whether the practice offered educational opportunities, and (E) whether the practice had a good relationship with hospitals. The authors recommended enhancing these factors in areas with labor shortages. Other practice factors, such as high deprivation allowance, opportunities to work privately, whether the practice was near where the trainee grew up, or whether the practice was where the trainee studied, were considered less important factors. A similar postal survey of 101 GP trainees in the south west region of England studied the factors that trainees viewed as important in making career decisions

(Rowsell, Morgan and Sarangi, 1995). The authors of that study found that time for leisure activities, on-call rota, maternity/paternity leave, and weekend work were the four most important factors in career decision-making. Less important factors included flexible work patterns, partner's paid work, salary, and regular work schedules. Yet another survey of 52 trainees located in North Trent found that trainees hoped to join "medium-sized practices with full ancillary and attached staff" and that relatively few were willing to work in an inner city or deprived area (Webb and Hannay, 1996).

One research team used responses to actual GP vacancy advertisements to assess which factors were most important to GPs (Carlisle and Johnstone, 1996). Specifically, the authors investigated the association between GP responses to 489 vacancies posted in the *British Medical Journal* between January and April 1995 and practice-level characteristics. They found that practices that did their own on-call work, practices not in the inner-city, and practices without deprived patients received a higher number of applications. The authors noted that, given their findings, it is possible that physicians may not have a strong preference for or against treating deprived patients but may instead simply wish to live in a non-deprived area.

Later studies used discrete choice models and conjoint analysis to determine GP's preferences for certain practice characteristics (**??**). For example, Scott (2001) conducted a discrete choice experiment among a sample of 848 British GPs, and found that the strongest determinant of preference was out-of-work hours. Wordsworth et al. (2004) similarly conducted a discrete choice experiment to assess which practice attributes were most important among GPs, giving special consideration to differences in the preferences among principal and sessional GPs. The authors found that, while some attributes had different levels of importance among principal and sessional GPs, both types of GPs preferred practices with less out-of-hours work and an increase in annual earnings. Gosden, Bowler and Sutton (2000) designed a questionnaire using conjoint analysis in which a respondent's preferences over eight characteristics could be derived from 27 practice scenarios. The respondents – a sample of 172 GPs who had recently taken on a principal post – were then asked to compare one hypothetical practice to each of the other 26 hypothetical practices. The authors found that the most important factor for GPs was aversion to moving to a high deprivation area, and that GPs preferred practices that had an extended primary healthcare team, afforded opportunities for the GPs to develop outside interests, paid higher salaries, had shorter working hours, and had shorter list sizes. Similar experiments have been used to elicit preferences from physicians in other develoed nations. For example, **?** performed a discrete choice experiment to discern what pecuniary and non-pecuniary job benefits would offset the disutility German physicians experience when moving from an urban to rural practice. The authors found that, all else equal, physicians required over 9,000 Euro per month to compensate for the disutility of moving from an urban to a rural practice, but that requirement diminished with the addition of nonpecuniary job benefits such as on-site childcare and fewer on-call duties. Collectively, these studies indicate that pay can be a major factor influencing whether a physician prefers one practice over another but that other non-financial incentives exist.

2.2 Other Forces for Movement

Only a few studies have considered what prompts a GP to want to move. (**?**) find that favorable wages and tax conditions induce Norwegian medical doctors to consider moving from other healthcare professions to full-time practice, which may indicate that financial incentives can motivate physicians to move.

Table 8: Neyman's Notation for a Completely Randomised Experiment

Unit	Potential Outcome		Individual-Level Causal Effect
	Treatment $Y(1)$	Control $Y(0)$	
1	$Y_1(1)$	$Y_1(0)$	$Y_1(1) - Y_1(0)$
...	...	...	...
i	$Y_i(1)$	$Y_i(0)$	$Y_i(1) - Y_i(0)$
...	...	...	...
N	$Y_N(1)$	$Y_N(0)$	$Y_N(1) - Y_N(0)$

NOTE: This table was adapted from Rubin (2005).

3 Methods for Evaluating the Effectiveness of Health Policies

3.1 Theory of Causal Inference

The core theory of causal inference, termed the "Potential Outcomes framework", comes from the largely separate work of two researchers in the early 20th century: Ronald Fisher (1890-1962) and Jerzy Neyman (1894-1981) (Rubin, 2005).

Neyman, then a Ph.D. student at the University of Warsaw, introduced the theory and, notably, a notational structure for causal inference in his dissertation. As shown in Table 8, individuals indexed at $i \in 1, ..., N$ would be observed in both the treatment and the control conditions. For each individual, the outcome in the control condition $Y_i(0)$ would be subtracted from the outcome of the same individual in the treatment condition $Y_i(0)$ to compute individual-level causal effects.

The sample Average Treatment Effect (ATE) would, by definition, be the mean of these individaul-level causal effects (1)

$$\text{ATE} = \sum_{i=1}^{N} \frac{Y_i(1) - Y_i(0)}{N} \tag{1}$$

Unfortunately, it is impossible to observe the same individual having been placed in the treatment condition and in the control condition — an individual can only live one existence. Neymand studied the ATE in the context of a simple, completely randomised experiment, i.e., a study in which participants are randomly assigned to either a treatment condition or a control condition. He found that, in this context, difference between the sample mean of those in the treatment condition and the sample mean of those in the control condition was an unbiased estimator of the Average Treatment Effect (2).

$$\mathbb{E}(\bar{Y_1} - \bar{Y_0}) = \mathbb{E}(\sum_{i=1}^{N} \frac{Y_i(1) - Y_i(0)}{N}) \tag{2}$$

That is, in the context of a completely randomised experiment, the difference in control group and treamtent group sample means, which can be observed, is an unbiased estimator of the true ATE, which cannot be observed. Neymand also described that the estimate of variance for a difference in sample means, $s_1^2/n_1 + s_0^2/n_0$, is either equal to or greater than the variance of the true ATE. This means that a significant difference between the control and experimental group sample means indicates a significant ATE.

Near the same time, Roland Fisher similarly conceived of a theory of cause and effect that relied upon knowing the outcome of a given individual had he received a treatment and had he not received a treatment. For example, Fischer stated in 1918 (Fisher, 1919): "If we say, 'This boy has grown tall because he has been well fed,' we are not merely tracing out the cause and effect in an individual instance; we are suggesting that he might quite probably have been worse fed, and that in this case he would have been shorter".

Fischer is generally credited as the founder of the randomised experiment (Rubin, 2005).[3] However, instead of treating randomisation in a theoretical conceptualisation, as Neyman had done, Fischer viewed randomisation as a way to test hypotheses in empirical work. In his 1926 textbook *Statistical Methods for Research Workers*, Fischer formally proposed the

[3] While Neyman's 1923 thesis used the concept of completely randomised experiments in his theory before Ronald Fischer proposed them in 1925 (Fisher, 1925). Neyman, himself, insisted that his theory relied on randomisation simply as a means to treat the outcomes probabilistically while Fischer demonstrated the necessity of randomisations in experiments beyond statistical implications (Reid, 1998).

"Fischer sharp null hypothesis" that $Y_i(0) = Y_i(0)$ $\forall i \in 1...N$. Under this formulation, the outcomes for any given individual are known for both the experimental and control condition, which is impossible. Instead a hypothesis test can be applied. If the Fischer sharp null hypothesis were true, $Y\bar{(}1) - Y\bar{(}0) = 0$, and a formal hypothesis test can be used to calculate the probability (i.e., p-value) that, given the observed data, y the observed difference between sample means $y\bar{(}1) - y\bar{(}0) = 0$.

Importantly, both Neyman's and Fischer's models depend upon a condition called the Stable Unit Treatment Value Assumption (SUTVA). Under SUTVA, an each individual's outcomes are independent; that is, assigning an individual to either the control condition or the treatment condition does not affect anyone else in the experiment. Further, the control and treatment conditions are assumed to be stable, regardless of the number of participants who are assigned to the condition. That is, an individual could expect the same experience regardless of whether many or few participants were assigned to her condition.

3.2 Causal Inference in Randomised Studies

It was not until the early 1970s that Rubin (1974) formally recognised and described the conceptual value of randomly assigning individuals to participate in the study and the conceptual value of randomly assigning participats to the control or treatment conditions. In a completely randomised experiment, an assignment vector $W = (W_1...W_N)$ is randomly generated such that each individual i in the population is assigned to either not participate in the study ($W_i = *$), participate in the study in the control condition ($W_i = 0$), or participate in the study in the treatment condition ($W_i = 1$). It is not necessary that the probability of assignment be the same (i.e., $P(W_i = *) = P(W_i = 0) = P(W_i = 1)$), so long as the probabilities are uniform for all individuals, i.e., $P(W_i = A) = P(W_j = A)$ $\forall i,j \in 1...N$ and $\forall A \in *,0,1$. In a randomised experiment, W exhibits two key features. The first, called "ignorability", is that the two potential outcomes for each participant (one observed and one unobserved) is independent of the randomisation (Rubin, 1974). The second is that each individual could be assigned to either the control or the treatment condition, i.e., $0 < P(W_i|X,Y(),Y(1)) < 1$.

In this case, the vector W is an entirely exogenous random variable; that is, no other factors influence the values that it takes on. Consequently, even accounting for a vector (or matrix) of external factors X, the vector of (observed or unobserved) outcome values given the treatment $Y(1)$ and vector of (observed or unobserved) potential outcome values given the control $Y(0)$, an unbiased estimator for the ATE can be computed in a completely randomised trial using Neyman's equality (2). This can be written formulaically as:

$$\mathbb{E}(\bar{y_1} - \bar{y_0}|X,Y(1),Y(0)) = \mathbb{E}(\bar{y_1} - \bar{y_0}) = \mathbb{E}(\sum_{i=1}^{N} \frac{Y_i(1) - Y_i(0)}{N}) \tag{3}$$

$$V(\bar{y_1} - \bar{y_0}|X,Y(1),Y(0)) = V(\bar{y_1} - \bar{y_0}) \geq \mathbb{E}(s_1^2/n_1 + s_0^2/n_0) \tag{4}$$

In sum, because the assignment vector W is randomly generated depending upon no other factors, it is trivial to compute an unbiased estimator of the causal effect and a positively biased estimator of the variance. Using these estimators, it is further possible to test the hypothesis that the effect is greater than a hypothesized value H (often $H = 0$).

3.3 Causal Inference in Non-Randomised Studies

By recognising the value of randomisation in experiments, Rubin (1974) was able to extend the Potential Outcomes Framework to account for non-randomised studies.

In non-randomised studies it is possible that the assignment vector W depends upon other factors. For example, a non-randomised study may show that people who drive sports cars are healthier than those who drive sedans. While there may be an effect of the car on health, it is also possible that an underlying third factor connects the treatment (i.e., car) with the outcome (i.e., health). Those who own sports cars may be wealthier than those who drive sedans, and wealth could be the causal driver of a difference in health.

In terms of the equations above, W is not independent with respect to X and so $\mathbb{E}(\bar{y_1} - \bar{y_0}|X,Y(0),Y(1)) \neq \mathbb{E}(\bar{y_1} - \bar{y_0})$. This source of bias is commonly referred to as "confounding". As another example, individuals might self-select into the treatment condition based upon their demographics, their circumstances, or their need for treatment. For example, it is known from randomised controlled trials that paracetamol reduces headaches. However, in an observational study of people using paracetamol and people not using paracetamol, it is likely that those using paracetamol are more likely to have a headache than those who are not. In this case, W is not independent with respect to $Y(0)$ and $Y(1)$ and so, again, $\mathbb{E}(\bar{y_1} - \bar{y_0}|X,Y(0),Y(1)) \neq \mathbb{E}(\bar{y_1} - \bar{y_0}$. This source of bias is commonly referred to as "self-selection". Further, in studies without random assignment, it is possible that the decision of whether to participate depends upon other factors. Researchers will only observe data from those who elected to participate, which could be unrepresentative of the population, and it is possible that the treatment may have a different effect if imposed upon those who were not included in the study. This may be referred to as "sample-selection" bias.

It is impossible to compute a truly assumption-free causal estimate without true random assignment. However, several methods have been designed to formally adjust for and minimise the restrictions levied by these assumptions. In this section, I will describe several common causal inference methods that can be used to generate causal estimates without random assignment, along with the assumptions that these models make.

3.3.1 Matching

A major source of concern in non-randomised studies is "self-selection" bias. Self-selection bias occurs when individuals choose, based upon thier own circumstances and preferences, whether to receive the treatment or control condition. In statistical terms, self-selection bias occurs when the assignment vector W depends upon endogenous variables. This form of bias is particularly dangerous, given that it is impossible for a researcher to know with certainty all the relevant endogenous variables within a system. That is, even if it could be confirmed that W does not depend upon any of the observed covariates in matrix X, it is possible that W depends upon one or more unobserved endogenous variables.

Matching is a common method to address self-selection (Stuart, 2010). In its simplest form, called exact matching, participants are matched into strata of others who share all observable endogenous variables (Imai, King and Stuart, 2008). Importantly, these matching variables should be selected from only those that could have influenced the individual's decision to receive the control or treatment condition; matching participants based upon ex-post variables (i.e., variables measured after the individual selected into a condition that could, hypothetically, be determined by their choice) is inappropriate. In this setting, participants within strata are identical in terms of the matching variables. Under the assumption that these matching variables constitute all the non-random factors that determine whether an individual chooses a given condition, then, assignment within a strata is effectively random. Consequently, a researcher can, within each strata, compare the sample means for individuals who received and did not receive the treatment to compute a within-strata treatment effect. To compute an ATE, a researcher takes the mean within-strata treatment effect weighted by the overall number of individuals within that strata (Rosenbaum and Rubin, 1984). While exact matching is straightforward, it has several disadvantages. Perhaps most importantly, only those observations with exact matches on all matching covariates can be included in the analysis. Further, only those strata which include both participants who received and did not received the intervention can be analysed – when treatment conditions is uniform within a strata, it is not possible to take the different in sample means. An analysis' precision is inversely proportional to the sample size, and so the sample size reductions of exact matching lead to imprecise estimates.

A closely related method to exact matching is coarsened exact matching (Iacus, King and Porro, 2009). With exact matching, individuals are grouped based upon whether they match exactly over all matching variables. Coarsened exact matching matching weakens this restriction, allowing for strata that have some differences in matching variables. First, a copy of the matching variables are made. Then, the copies are "coarsened" such that like, but not necessarily the same, values of each variable are grouped together. The extent to which like values are grouped can either be set by the researcher or determined algorithmically (Blackwell et al., 2009). Individuals are then placed into strata based upon having the same coarsened matching variables. The coarsened exact matching model, consequently, allows the researcher to use a greater proportion of her data; those individuals with matching variable values that are similar to but not exactly the same as others can still be analysed after coarsened exact matching. However, the assumption persists that the matching variables (in this case, the coarsened matching variables) explain all non-random variation in which condition the participant chose.

While the exact and coarsened exact matching methods are straightforward, they are relatively uncommon in the literature. The most commonly implemented form of matching is called "Propensity Score Matching" (Rosenbaum and Rubin, 1983). Similar to the previous two models, Propensity Score Matching relies on the notion that observed covariates explain which condition an individual chooses, up to random variation (Rosenbaum and Rubin, 1985). Unlike the other two models, however, Propensity Score Matching reduces all these matching covariates into a single vector, which represents the probability that an individual will be treated conditional upon the matching covariates. That is, an individual i would be given the propensity score $e_i = P(T_i = 1|X_i)$. Probability is commonly calculated through logistic regression or non-paramtric boosted methods (Stuart, 2010). Notably, the distribution of matching covariates are the same between individuals in the treatment and individuals in the control condition at each discrete propensity score value, and the distribution is similar between the two is similar among similar propensity score values. Consequently, a researcher can estimate treatment effects within strata of identical propensity scores (as in exact matching) or, more commonly, within strata of similar propensity scores and then aggregate them to compute an ATE. To group together observations of similar propensity scores, first, the distance is calculated between every possible pair of individuals such that disttance between individual i and individual j is $D_{ij} = |e_i - e_j|$ (the linear propensity score distance formula $D_{ij} = |\text{logit}(e_i) - \text{logit}(e_j)|$ can also be used; (Rosenbaum and Rubin, 1985)). Then, an algorithm can be used to match pairs with similar propensity scores but dissimilar conditions. For example, the 1:1 nearest neighbor algorithm pairs each treated individual with the least distant control individual (Stuart, 2010).

3.3.2 Regression Modelling

Perhaps the most commonly used modern statistical technique is regression. While not inherently a causal inference method, regression can be used to mitigate confounding bias in non-randomised studies and forms the foundation for several significant causal inference methods (Zohoori and Savitz, 1997).

The simplest form of regression is Ordinary Least Squares Regression (OLS). Other regression models exist to account for different distributions of the data (e.g., logistic or probit regression can be used for binary outcome data, Poisson or negative binomial regression can be used for count outcome data, beta regression can be used for outcome variables lie between 0 and 1), but the use of regression to account for confounding is similar. The model underlying OLS is shown as:

$$Y = X\beta + \epsilon \tag{5}$$

In this model, Y is a $n \times 1$ observable matrix of all outcome values, X is a $n \times k$ observable matrix constructed by binding columns of k independent variables together, and ϵ is a $n \times 1$ matrix of all errors. β represents the $k \times 1$ matrix of non-random but unobserved parameters that minimise the function $\sum_{i=1}^{n} \epsilon^2$. When expectation of the errors are 0, that errors have common variance (homoskadastic), and that errors are uncorrelated, it can be shown that the OLS estimators $\hat{\beta} = (X'X)^{-1}X'y$ are also lowest-variance unbiased estimators of the true coefficients β (Puntanen and Styan, 1989).

In a completely randomised experiment with Y as the outcome and X as an indicator variable equal to 1 if the individual was in the treatment condition and 0 if the individual was in the control condition, these assumptions on the error terms would likely be met. The distribution of covariates would be identical between those in the control and treatment conditions, and so any other factor that has an influence on the outcome would be evenly distributed between the control and treatment conditions.

However, in non-randomised studies, it is likely that the distribution of covariates would be unequal between groups, and the parameter estimate may be biased by confounders disproportionate to certain groups. Fortunately, if, *conditional on all other independent variables*, the expectation of the error term is 0, the errors have common variance, and the errors are uncorrelated, the OLS estimators will still be the lowest-variance and unbiased estimators. Consequently, it is possible to reduce confounding bias by including variables for potential confounders in the regression model.

A crucial assumption for this method is that all potential confounders are included in the model (Crémieux and Ouellette, 2001). It is relatively simple to include observed confounders when computing the model; however, unobserved confounders can also bias the OLS estimators and, as the researcher cannot observe them, cannot be accounted for in regression. This form of bias is commonly referred to as “omitted variable bias”; without random assignment, it is impossible to entirely eliminate the possibility of omitted variable bias.

3.3.3 Multi-Level Models

While it is impossible to entirely eliminate the possibility of omitted variable bias, bias presented by certain types of unobserved confounders can be excluded when data is nested. In many empirical settings, observations are naturally grouped together (Duncan, Jones and Moon, 1996). For example, in longitudinal data, individuals may complete a survey at several different time points. Each individual-by-time observation then fits within a group of same-individual observations (i.e., all the observations from that same individual over all time points) and within a group of same-time observations (i.e., all observations from that same time point over all individuals). For another example, individual-level panel data may be grouped by jurisdiction (e.g., state) and by time, even if the individuals contributing data are different from year to year. In these settings with nested data, fixed and random effects can be used to account for confounding that is invariant within groups, even when the confounders are unknown (Leyland and Groenewegen, 2003).

The most basic tool for multi-level models is fixed effects. Consider, for example, a study aiming to assess whether a procedure changes length of hospital stay using deidentified, patient-level data on procedures from several hospitals. Observations are nested within hospitals, such that each hospital contains some some observations where the procedure was performed and some observations where the procedure was not performed. A naiive regression would simply compare the length of hospital stay among those who had the procedure versus those who did not (equation (6)).

$$y_i = \beta_0 + \beta_1 * (\text{Procedure} = 1) + \epsilon_i \tag{6}$$

However, there are likely several variables that confound the relationship between the exposure (receiving the procedure) and the outcome (length of hospital stay). Some of these confounders may vary at the individual level; for example, if increased severity of the condition both causes the physician to use the procedure and leads to increased length of hospital stay after the procedure, then severity of the condition would be a confounder. If patient-level data on condition severity is available, it should, consequently, be added to the regression model. If data is not available for condition severity, the regression cannot be adequately adjusted, which could lead to omitted veriable bias. While all confounders have at least some variance at the individual-level, some confounders may vary predominantly at the hospital level. For example, it is possible that some hospitals serve wealthier populations than others, which could influence both the exposure and the outcome. Even if individual-level data is unavailable for the wealth of the patient, accounting for the hospital through a hospital fixed effect could account for the hospital-level variation in wealth of patients. Further, a hospital fixed effect accounts for hospital-level variation in all potential confounders, including both observed and unobserved confounding. That is, if the researchers fail to consider an important confounder but that confounder varies primarily at the hospital level, then a hospital fixed effect may, at least partially, reduce the bias from that variable. Conceptually, the fixed effect groups together observations within the same hospital and estimates the effect of the exposure on the outcome within each hospital. The final effect estimate, then, is the weighted average of the effect of the procedure within each hospital.

A simple fixed effect, like that described in the previous example, is implemented by setting a separate baseline length of stay (i.e., mean length of stay for participants who did not receive the first procedure) for each group j:

$$y_{ij} = \beta_1 * (\text{Procedure} = 1) + \alpha_j + \epsilon_{ij} \tag{7}$$

A common technique using fixed effects in regression is intercept fixed effects (Bell, Fairbrother and Jones, 2019). Intercept fixed effects essentially set a a different intercept for the outcome variable within each group. (Equivalently and often more efficiently, fixed effects can be implemented by subtracting the group-level mean from each individual in each group). In this way, the fixed effects account for baseline differences among the groups of data. These fixed effects also account for any confounder that does not vary with the independent variable within the group. That is, if a given confounder is invariant within groups (e.g., sex of a participant or location of a hospital), then that effect on the outcome will be present at the intercept. Consequently, using a separate intercept for each group will eliminate the effect of that confounder. Intercept fixed effect correspond graphically to parallel, group-level lines on either side of the weighted average intercept model (Figure 1, Panel B). The slope of these lines is a pooled estimate, and so it is the same among all groups. The intercept, however, is individually determined, and so the lines may be parallel.

A common implementation of intercept fixed effects uses panel data, where data is collected for each individual i within several groups j over several years t, and a separate intercept is set for each group and year:

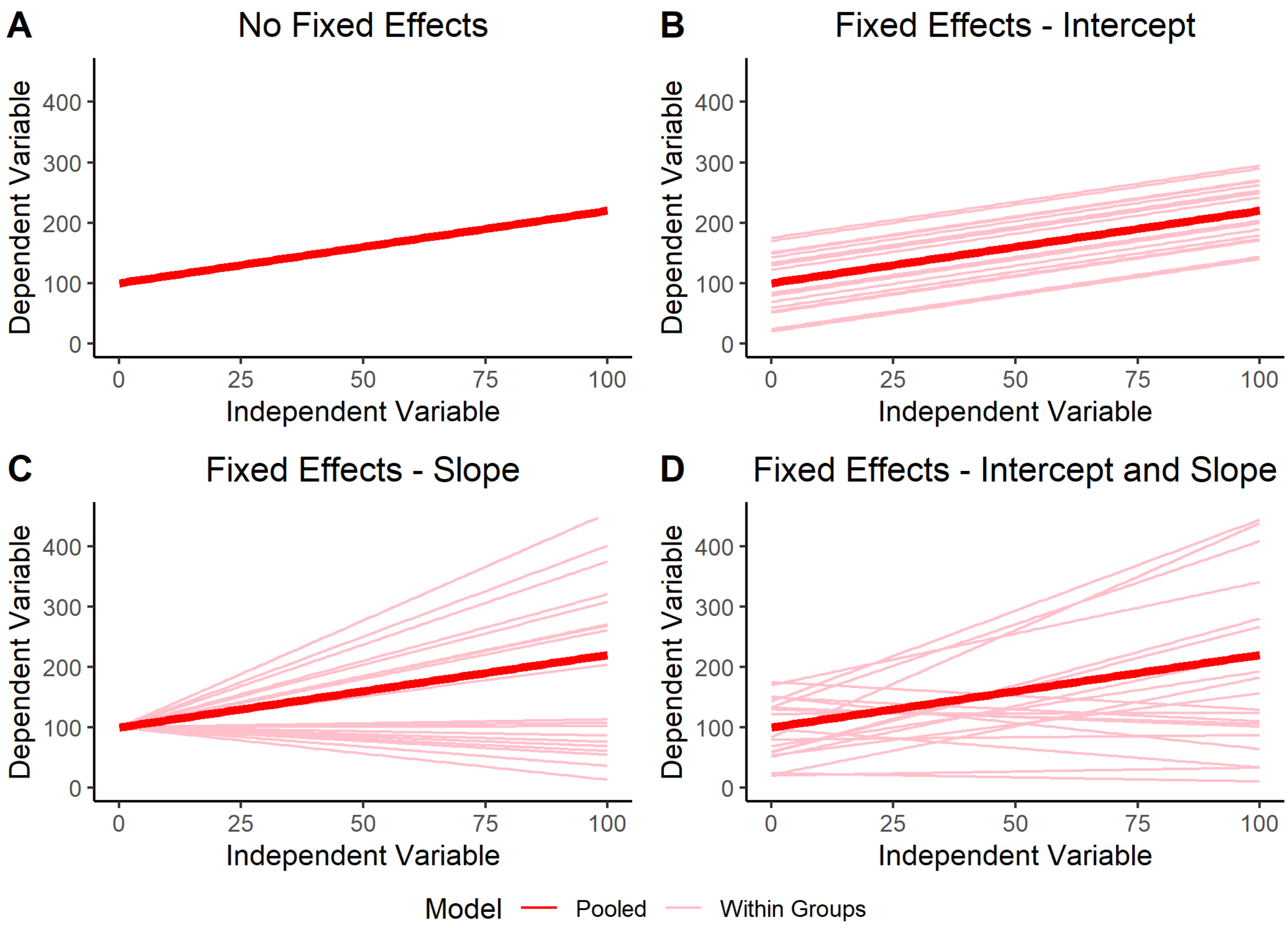
$$y_{ijt} = \beta_1 * \text{Exposure} + \alpha_j + \lambda_t + \epsilon_{ijt} \tag{8}$$

In this model, a separate intercept is set for each group, and each observation's outcome is measured in relation to the mean outcome within that time. The group fixed effect accounts for time-invariant confounding at the group level, while the time fixed effect accounts for group-invariant confounding at the year level. Notably, group-level fixed effects cannot account for confounders that vary within the group; this model does not account for confounder variation for individuals within groups or for confounders that change over time within groups.

While intercept fixed effects are perhaps the most common, the concept of fixed effects can be applied to any parameter in a regression model, as long as the model has sufficient degrees of freedom (i.e., the model matrix is not overdetermined). For example, slope fixed effects assume that the relationship between an independent variable and the outcome is constant within groups. Slope fixed effects can be implemented by interacting the relevant independent variable with the group-level fixed effects. When using fixed slopes, the intercept is a pooled estimate that is constant across all groups, while the slopes are estimated independently. Consequently, a fixed slopes model corresponds graphically to a single origin point with several group-level lines that are steeper or less steep than the weighted average slope model (Figure 1, Panel C).

Further, in cases where there is sufficient data (and degrees of freedom), fixed effects can be used for both the intercept and the slope within the same model. Fixed intercept effects allow for the intercept to vary between groups, and fixed slope effects allow for the slope to vary between groups. Using both fixed intercept and fixed slope effects, both the intercept and the slope can vary between groups. Graphically, this corresponds to several non-parallel lines not emerging from a single point (Figure 1, Panel D). Depending upon

Figure 1: Illustration of Fixed Effects



NOTE: These figures were generated from simulations by the author. Group intercepts take on values randomly generated between 20 and 180, and group slopes take on values between -1.2 and 3.6.

the between-group variance in the intercept and the slope, these lines can be similar to or different from the model for the weighted average slope and intercept model.

An important feature of fixed effects is that they carry the assumption that the groups are truly independent with respect to the parameter being estimated. On one hand, this can be a valuable feature; a researcher does not need to assume a certain group-level distribution of the underlying parameter when using fixed effects. Because fixed effects are estimated entirely within groups, they can appropriately account for group-invariant confounding even if the impact of group on the parameter is highly heterogeneous. On the other hand, this feature limits the interpretability of any parameter estimated through fixed effects. Because parameters with fixed effects are estimated entirely within groups, the parameter estimate is only meaningful within that group and cannot be extrapolated to other groups. For this reason, fixed effects are typically used to rule out confounding by lurking variables. If fixed effects are applied to the independent variable of interest, the parameter estimates for that variable will be uninterpretable.

One major disadvantage of fixed effects is their inefficiency. Parameter estimates become more imprecise with decreasing degrees of freedom, and including a fixed effect reduces a model's degrees of freedom by the number of levels it takes on. In this respect, the related concept of random effects provide a distinct advantage. Random effects account for between-group variation in confounders with a much lesser impact on the model's efficiency. In exchange, however, random effects carry some assumptions on the group-level distribution of the parameter.

Like with fixed effects, the purpose of random effects is to adjust for observed and unobserved group-level variation in naturally nested data. Unlike fixed effects, however, the random effects method assumes that the group-level parameter estimates come from a known, typically normal distribution (Puntanen and Styan, 1989). In this way, the group-level parameter estimates are not strictly independent and can instead be modelled as a random variable. Consequently, these group-level parameters are estimated through "partial pooling", accounting to some extent for the distribution of the other group-level parameters, and group-level parameter estimates that vary too far from the distribution are pulled towards the mean.

A simple random effects model is shown in equation (9). In this model, y represents a vector of $N \times 1$ matrix containing the values of the dependent variable, Z represents a $N \times q$ "design matrix" where q is the number of groups in the data, u is a $q \times 1$ matrix of random effects, and ϵ is a $N \times 1$ matrix of the errors.

$$y = \beta_0 + Zu + \epsilon \tag{9}$$

The design matrix Z is constructed such that each row represents an individual observation and each column represents a group. When observation i is in group j, $Z_{ij} = 1$. Otherwise, $Z_{ij} = 0$. This is a familiar construction; Z is simply a collection of dummy variables for whether the observation fits within a given group. Indeed, if the parameters in u were estimated through OLS, then u would be a vector of group-level fixed effects. However, as previously noted, random effects impose some assumptions on the estimators, distinguishing them from fixed effects. Instead of directly estimating u, it is assumed that $u \sim N(0, G)$. u is distributed around 0 because the intercept β_0 absorbs the pooled mean for y, and G is the variance-covariance matrix of u. In this case, the intercept is the only parameter estimated through random effects and u is a single-variable, $q \times 1$ matrix, so the variance-covariance matrix of u is simply the variance of the random intercept. In cases where more than one parameter is estimated by the random effects model, it may be necessary to place additional constraints on the covariance structure among random effects (e.g., independence) to efficiently estimate G.

In more concrete terms, a random effects model can be illustrated in a regression model, as in equation (10).

$$y_{ij} = (\gamma_{00} + u_{0j}) + \epsilon_{ij} \tag{10}$$

In this regression, y_{ij} represents the outcome for observation i in group j, γ_{00} represents a common intercept for all observations, u_{0j} is a random variable centered at 0 with variance G that takes on a different value for each group j, and ϵ_{ij} is the random error term. Collectively, $(\gamma_{00} + u_{0j})$ is the group-level intercept, centered at γ_{00}.

Like fixed effects, the random effects method can be applied to any parameter in a regression – not just the intercept. Consequently, the graphic representation in Figure 1 is roughly applicable to random effects. When random effects are applied to the intercept, group-level estimates will parallel the weighted average model. When random effects are applied to the slope, group-level esitmates will emerge from the same intercept but have span outwards from the weighted average model. When random effects are applied to both the intercept and the slope, group-level estimates need not have anything in common with the weighted average model.

In sum, in cases where it is important to maximise the efficiency of a model and it can reasonably be assumed that group-level parameters are drawn from a random distribution, random effects are a suitable solution to addressing between-group confounding.

In many cases, it is appropriate to use both fixed and random effects. These models – referred to as mixed models – apply fixed effects and random effects methods to different parameters in the regression. For example, a regression could use intercept fixed effects and slope random effects. Mixed methods are particualrly useful when reasonable assumptions can be made on the distribution of some group-level parameters but not others. A simple mixed model can be written using equation (11).

$$y = X\beta + Zu + \epsilon \tag{11}$$

In this model, y is the $N \times 1$ matrix of outcomes, X is a $N \times p$ matrix, where p is the number of fixed effects to be included in the model, β is a $p \times 1$ matrix of fixed effects parameter estimates, Z is a $N \times q$ matrix where q is the number of groups used for random effects, u is a $q \times 1$ matrix for the random effects estimates, and ϵ is a $N \times 1$ matrix of errors.

3.3.4 Difference-in-Difference

One of the most common methods for causal inference is the difference-in-difference design (Lechner, 2010). Recall that in a completely controlled experiment, it is appropriate to merely compare the sample mean outcome value between those who received the treatment $Y(1)$ and those who did not receive the treatment $Y(0)$ to estimate a causal effect. The Difference-in-Difference design creates a parallel to this reasoning for non-randomised studies.

In its simplest possible form, a researcher wishing to use difference-in-difference attempts to find two samples of individuals that are highly alike on all observable characteristics

Table 9: Two Groups Identical at Baseline

	Outcome$_{t=1}$	**Outcome**$_{t=2}$
Control $Y(0)$	10	20
Treatment $Y(1)$	10	40

NOTE: This table was created by the author. Notice that each group have the same outcome value at $t = 1$. This is important, as it was assumed that the two groups were highly alike.

(including the outcome of interest) at baseline (i.e., before the intervention), with one sample experiencing an intervention and the other not experiencing the intervention. Under the assumption that the two groups were highly similar at baseline and extrapolating that the two groups would have experienced the same trajectory in the outcome variable had neither group experienced the intervention, the difference between the two groups after the intervention could be considered an estimate of the causal effect of the intervention. As an example (Table 9), the difference between the dependent variable in the treatment group and the control group ($40 - 20 = 20$) could be considered a causal effect of the treatment.

The assumption that the two groups were virtually identical at baseline is very strong. In virtually all non-randomised settings, there will be some baseline differences between the treatment and control samples. The difference-in-difference approach accounts for these issues using two differencing methods to construct a counterfactuals. The first difference occurs between the outcome values for the treatment group at baseline and the control group at baseline, accounting for differences at baseline. In this way, the control group becomes a more realistic, comparable counterfactual for the treatment group. The second difference occurs between the outcome values for the treatment group after the intervention and before the intervention. This difference allows the treatment group's pre-intervention outcome values to serve as a counterfactual for the treatment group's post-intervention outcome values. In sum, the difference-in-difference approach estimates the causal effect of the treatment as shown in equation (12), where y_{11} is the sample mean outcome for the treatment group in period 1, y_{21} is the sample mean outcome for the control group in period 1, y_{12} is the sample

Table 10: Simple Difference in Difference

	Outcome$_{t=1}$	**Outcome**$_{t=2}$	**Changes**
Control $Y(0)$	$y_{21} = 30$	$y_{22} = 25$	$y_{21} - y_{22} = 5$
Treatment $Y(1)$	$y_{11} = 60$	$y_{12} = 40$	$y_{11} - y_{12} = 20$
Difference	$y_{11} - y_{21} = 30$	$y_{12} - y_{22} = 15$	$(y_{12} - y_{22}) - (y_{11} - y_{21}) = -15$

NOTE: This table was created by the author.

mean outcome for the treatment group in period 2, and y_{22} is the sample mean outcome for the control group in period 2.

$$\text{DD} = (y_{12} - y_{22}) - (y_{11} - y_{21}) \tag{12}$$

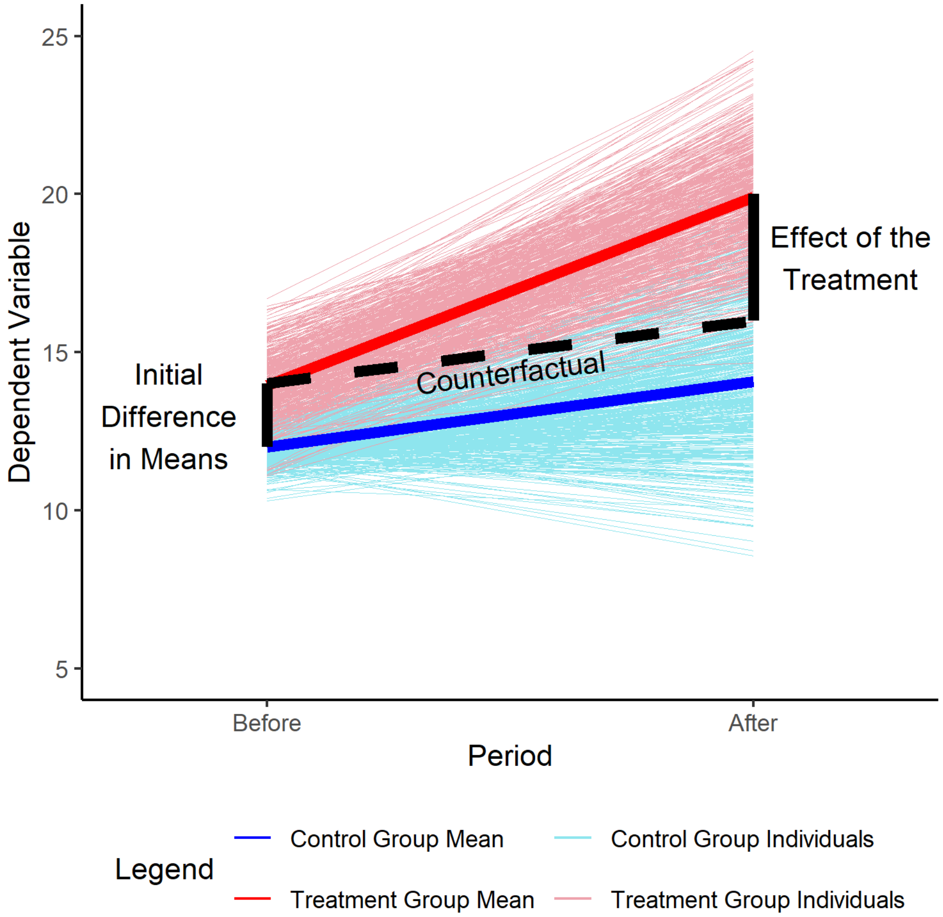
Table 10 shows an example of how the difference-in-difference estimator is calculated. The difference-in-difference estimator from this example is $(y_{12} - y_{22}) - (y_{11} - y_{21}) = -15$.

The causal effect from the difference-in-difference method can be shown graphically (Figure 2). From this representation, it is shown that the trajectory of the control group combined with the baseline difference between the treatment and control group create a counterfactual. The causal effect is illustrated by the black, vertical bar on the right of the figure, which highlights the sample mean for the treatment group in excess of what would be predicted by the counterfactual. This surplus of the dependent variable can be considered the causal effect of the treatment.

While this simple Difference-in-Difference model accounts for baseline sample mean differences in the outcome variable between the treatment and control group, it is still likely that the treatment and control group vary on other observable confounders. The Difference-in-Difference model can be extended to further account for confounding using regression, as shown in equation (13).[4]

[4] For an example of how this method can be used in practice alongside propensity score matching, see (**?**)

Figure 2: Illustration of Difference-in-Difference Design



NOTE: These figures were generated from simulations by the author. The control group had a mean of 12 before the intervention and an increase of 2 (SD=2) by period 2. The treatment group had a mean of 14 before the intervention and an increase of 6 (SD=1.5) by period 2. The causal effect can be calculated as DD = (20 − 14) − (14 − 12) = 4, which is approximately the height of the bar on the right side of the figure.

$$y_{ist} = \beta_0 + \beta_1 * \text{Period} + \beta_2 * \text{Condition} + \beta_3 * \text{Period} * \text{Condition} + X_{ist}\beta + \epsilon_{ist} \tag{13}$$

In this model, y_{ist} refers to the outcome value for individual i in condition s during period t, period is an indicator variable equal to 0 before the intervention and 1 after the intervention, condition is an indicator variable equal to 0 for the control group and 1 for the treatment group, and X_{ist} is a $N \times p$ matrix constructed by binding p confounder variables by column. β_0 is the parameter for the baseline outcome value among the control group before the intervention. β_2 is the paramater for the difference between the baseline outcome value for the treatment and the control groups. β_3 is the parameter of interest, corresponding with the difference-in-difference estimator. β is a column vector corresponding to paramater estimates for the p confounders, and ϵ_{ist} is the random error term. When the same individuals are sampled in each group before and after the intervention, the regression model can also account for individual-level fixed intercept effects (see Section 3.3.3), which essentially transforms the outcome to be the individual-level difference in the outcome before and after the intervention.

The underlying assumption for all Difference-in-Difference models is that, absent the intervention, the treatment group would have had the same trajectory as the control group in terms of the outcome variable. This is a strong assumption. Choosing samples that are similar at baseline and using regression to adjust for observable confounders somewhat mitigates this concern, though it is still possible that unobserved confounders could bias the difference-in-difference estimator.

3.3.5 Interrupted Time Series Analysis

Like Difference-in-Difference, Interrupted Time Series Analysis (ITSA) uses differencing to construct a counterfactual. There are two main forms of ITSA. One-sample ITSA uses the trajectory of the outcome within a group *before* an intervention to construct a counterfactual, against which post-intervention outcome values can be compared. This is similar to the first difference in the difference-in-difference approach ($y_{11} - y_{21}$). Two-sample ITSA uses both the trajectory of the treatment group before the intervention and the trajectory of an untreated control group to construct a counterfactual, corresponding with the full difference-in-difference model ($(y_{12} - y_{22}) - (y_{11} - y_{21})$). Both differ from difference-in-difference, however, in their treatment of time. In ITSA, the outcome variable is observed at several time periods before the intervention and at least one period after the intervention (Craig et al., 2012). Consequently, it is possible to observe not just the level of the outcome variable in each group before the intervention but also the trends over time before the interruption.

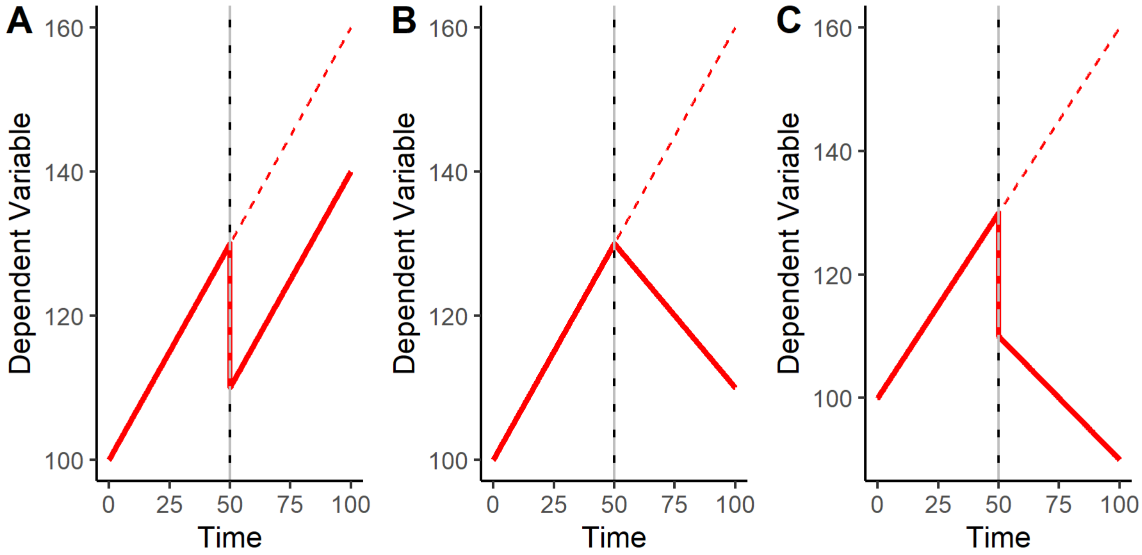
Because ITSA accounts for several time periods, it is almost always computed using regression. The regression model for a one-sample ITSA is shown in equation (14) (Lopez Bernal, Cummins and Gasparrini, 2016).

$$y_{it} = \beta_0 + \beta_1 * \text{Time} + \beta_2 * \text{Post} + \beta_3 * \text{Time} * \text{Post} + \epsilon_{it} \tag{14}$$

In this model, y_{it} is the outcome for individual i in time t, Time is a linear time term (e.g., 0 in period 1, 1 in period 2, etc.), and Post is an indicator variable if the observation is after the intervention. β_0 is the parameter for the intercept (the average outcome at period 1) and β_1 is the paramater for the slope on time before the interruption. Notably, as opposed to the difference-in-difference model which only reports mean differences, the interrupted time series analysis model decomposes the effect into an immediate effect and a gradual effect: β_3 corresponds to the change in level after the intervention or the immediate effect of the intervention. β_4 corresponds to the change in slope after the interruption or the gradual efect of the intervention. ϵ_{it} refers to the random errors in the model.

The effects represented by β_2 and β_3 can be visualised graphically. Panel A of Figure 3 shows an immediate effect (β_2) on the dependent variable following the intervention. Panel B shows a change in slope at the point of the interruption, representing a gradual effect on

Figure 3: Illustration of One Sample Interrupted Time Series



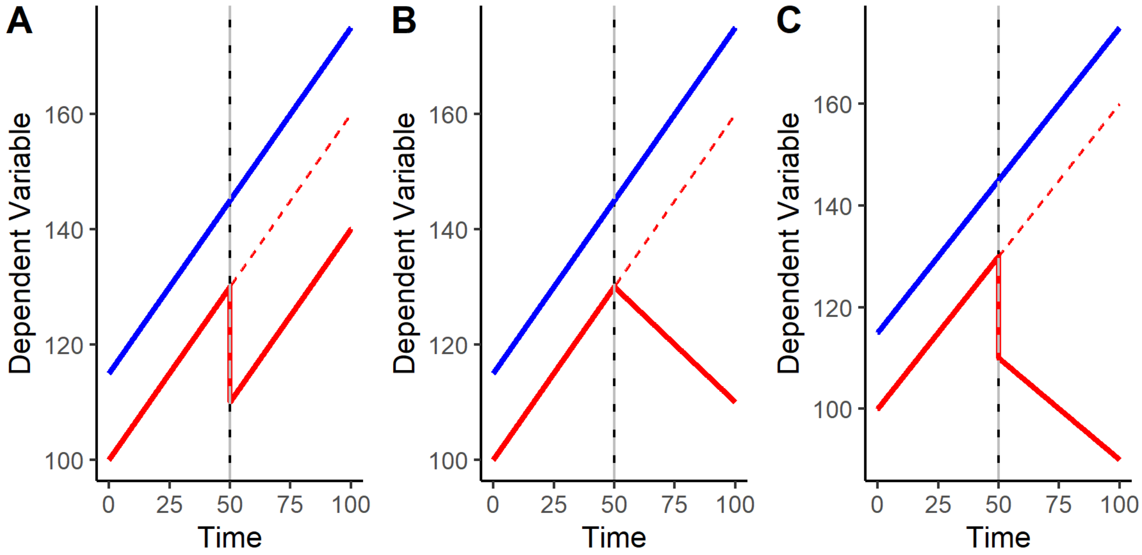
NOTE: These figures were generated from simulations by the author. The dotted red line in each panel shows the counterfactual, while the solid red line presents observed values for the sample studied. The interruption is set at time 50, noted with a dashed gray line. The total, observation-level effect of the intervention over the entire study period is represented by the space between the counterfactual (dashed red line) and the observed time series (solid red line) after the intervention. The initial intercept is set at 100, and the initial slope is set at 0.6. The level change is set to 20, and the post-intervention slope is set to -0.4.

the dependent variable (β_3). Panel C shows both these effects in the same time series.

The underlying assumption of the one-sample ITSA is that, in the absence of the intervention, the trajectory would have continued with the initial slope and intercept (Penfold and Zhang, 2013). In cases where the trend before the intervention – sometimes called the pre-trend – is consistent, precise, and linear and there are no other confounding events occurring near the time of the intervention, this may be a reasonable assumption. However, if the pretrend is not consistent or linear or if there are other confounding events occurring at the time of the intervention, it can be difficult to convincingly argue that this assumption would hold, and causal estimates from the one-sample ITSA may be biased.

The two-sample ITSA provides the researcher with an opportunity to empirically argue this assumption (Lopez Bernal, Cummins and Gasparrini, 2018). Like the difference-in-difference model, the two-sample ITSA uses a control group that was not affected by the intervention, as well as the treatment group pretrend, to construct a counterfactual. Conse-

Figure 4: Illustration of Two Sample Interrupted Time Series



NOTE: These figures were generated from simulations by the author. The dotted red line in each panel shows the counterfactual, while the solid red line presents observed values for the treatment group and the solid blue line represents observed values for the control group. The interruption is set at time 50, noted with a dashed gray line. The total, observation-level effect of the intervention over the entire study period is represented by the space between the counterfactual (dashed red line) and the observed time series (solid red line) after the intervention. For the treatment group, The initial intercept is set at 100, and the initial slope is set at 0.6. The level change is set to 20, and the post-intervention slope is set to -0.4.

quently, the control group provides the researcher with an opportunity to demonstrate that, absent the intervention, the trajectory would have continued from the pre-trend. That is, if the researcher can show that the control group and the treatment group have parallel pre-trends and that, after the intervention, the control group's trajectory follows the pretrend, it is often reasonable to assume that, absent the intervention, the treatment group also would have followed the same pretrend. The assumption, then, is that the control group and the treatment group would have had parallel trends throughout the study period in the absence of the intervention. This is commonly called the "parallel trends assumption". The design of the two-sample ITSA is shown in Figure 4.

The regression for two-sample ITSA follows the same basic form as that for one-sample ITSA (14) but with each parameter interacted with condition, as shown in equation (15).

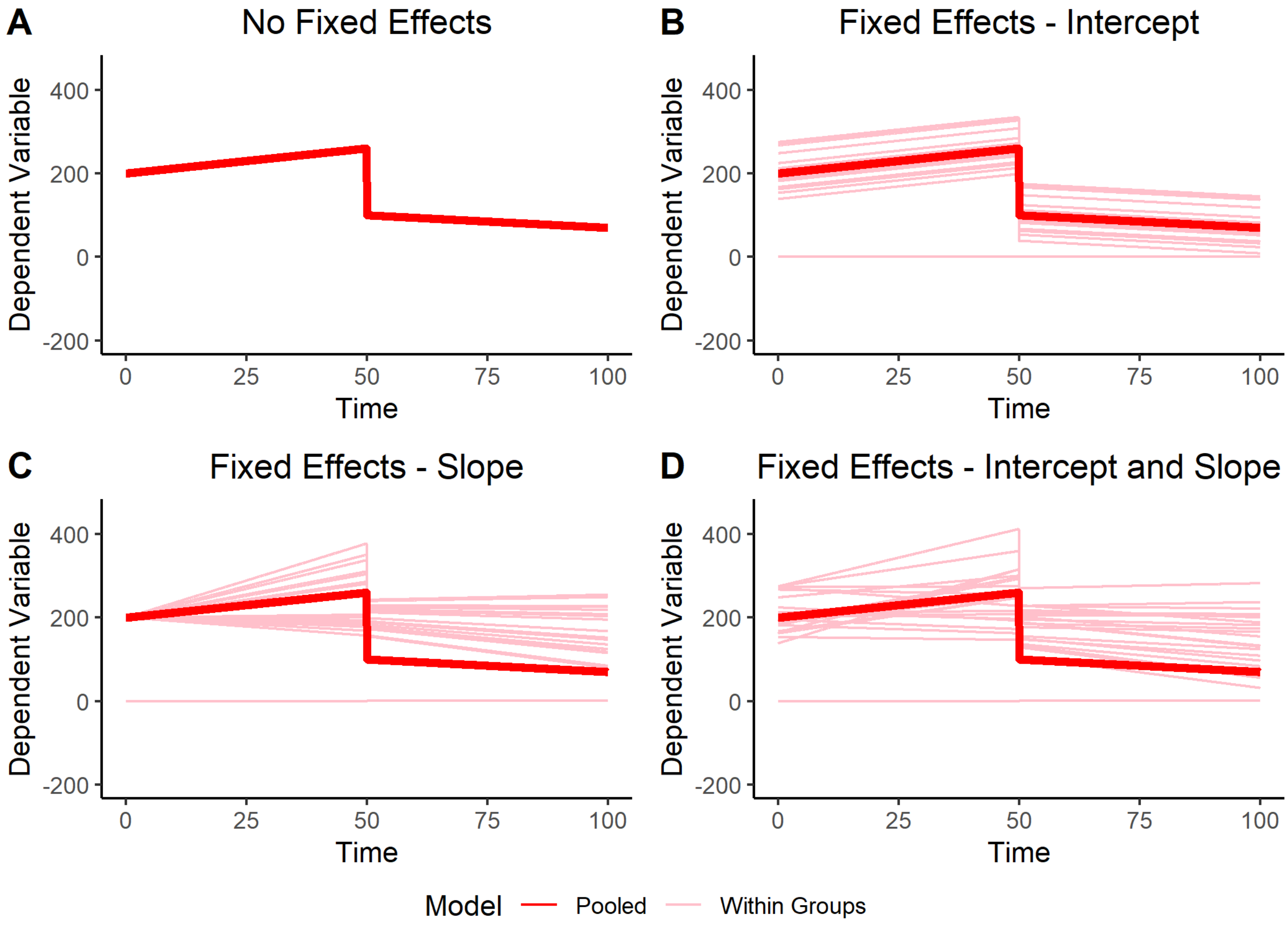
$$
\begin{aligned}
y_{ist} = \beta_0 &+ \beta_1 * \text{Time} + \beta_2 * \text{Post} + \beta_3 * \text{Condition} \\
&+ \beta_4 * \text{Time} * \text{Post} + \beta_5 * \text{Time} * \text{Condition} \\
&+ \beta_6 * \text{Post} * \text{Condition} + \beta_7 * \text{Time} * \text{Post} * \text{Condition} \\
&+ \epsilon_{it}
\end{aligned} \tag{15}
$$

In this model, y_{ist} is the outcome for individual i in condition s at time t, Time and Post are defined as in equation (14), and Condition is an indicator variable equal to 0 if the individual is in the control group and 1 if the individual is in the treatment group. β_0, β_1, β_2, and β_4 are parameter estimates corresponding to the intercept, initial slope with time, post-intervention level change, and post-intervention slope change for the control group. β_3, β_5, β_6, and β_7 are parameter estimates corresponding to the difference in intercept, difference in initial slope with time, difference in post-intervention level change, and difference in post-intervention slope change for the treatment group, relative to the control group. The parameters of interest are β_6, which corresponds to the immediate effect of the intervention, and β_7, which corresponds to the gradual effect of the intervention. Parallel pre-trends can be empirically tested through the estimate for β_4; if the estimate for β_4 is significantly different from 0 (usually with an $\alpha = 0.10$), then the pre-trends are not parallel between the control and treatment group. The estimate for β_3, while less important than parallel pre-trends, is often also reported to demonstrate that the control and treatment group had significant values of the outcome at baseline.

While the two-sample ITSA is a stronger design than the one-sample ITSA oweing to the use of a contemporaneous control group, it still has several limitations. While strongly parallel pre-trends signals (and is a prequisite of) the parallel trends assumption, it is not conclusive, and the researcher is forced to make some assumptions about what the trend of the treatment group would have been if those individuals had not experienced the intervention. Some of these concerns may be addressed by adding confounders into the regression model, though the possibility of unobserved confounding and omitted variable bias remains. Further, both one-sample and two-sample ITSA assume that there were no other relevant interventions occuring in either of the two groups at around the same time as the intervention. If two relevant interventions occur at nearly the same time, it is nearly impossible to disentangle their effects through an ITSA.

An alternative method to weakening the assumptions imposed by the one-sample ITSA is through the use of fixed and random effects. ITSA often uses data from individuals collected over the course of the study period (i.e., before and after the study period). In this case, the individual-by-time observations are naturally grouped by individual, and individual-level fixed or random effects can be applied when estimating the ITSA parameters. As discussed in Section 3.3.3, fixed effects or random effects can be applied to any of the parameters in the regression (though it does not make sense to apply fixed effects or random effects to the parameters of interest). Applying fixed effects or random effects to different parameters weakens some underlying assumptions of ITSA and allows the model to more precisely estimate the effect of the intervention on a given individual. For example, adding intercept fixed effects to the ITSA regression removes the assumption that all individuals experience the same immediate effect of the treatment and allows the pooled ITSA immediate effect estimate to account for within-individual immediate effects of the intervention. Fixed effects in ITSA is shown graphically in Figure 5, Panel B. Fixed slope effects, shown graphically in Panel C, remove the assumption that all individuals experience the same change in slope before and after the treatment, allowing the ITSA gradual effect estimate to account for within-individual changes in trajectroy before and after the treatment. Fixed intercept and slope effects, as shown in Panel D, remove both these assumptions. Using random intercept and slope effects have largely the same goals as using fixed intercept and slope effects, though with better efficiency and additional assumptions on thhe underlying distribution of the parameter estimates (see Section 3.3.3). For an example using fixed and random effects

Figure 5: Illustration of Interrupted Time Series Analysis with Fixed Effects



NOTE: These figures were generated from simulations by the author. Group intercepts take on values randomly generated between 120 and 280. Initial slopes take on values between -1.2 and 3.6. Post-interruption intercepts take on values between 180 and 280. Post-interruption slopes take on values between -1.5 and 0.3.

in a two-sample interrupted time series analysis, see (**?**).

3.3.6 Synthetic Controls

In some cases, a researcher wishing to use the two-sample ITSA method has access to data from several possible control groups, but no control groups demonstrate parallel pre-trends. In the ITSA framework, using a control group with non-parallel pre-trends would be severely damaging to the internal validity of the study, and ITSA estimates would likely be biased. To overcome this concern, the researcher could use a modern extension of the ITSA framework

Table 11: Illustration of Synthetic Controls

Time	C1	C2	C3	Treatment
0	2	2	-1	1
1	4	0	2	2
2	0	6	-1	4
Intervention				
3	3	-2	0	6
4	1	4	2	7

NOTE: This table was created by the author.

called "Synthetic Controls" (Craig, 2015).

Conceptually, the Synthetic Controls method uses a weighted sum of several different control groups in order to construct a control group with nearly identical pretrends to the treatment group (Abadie, Diamond and Hainmueller, 2015). For a simple example, shown in Table 11, no possible control group (C1, C2, or C3) has a parallel pre-trend to the Treatment group.

However, it can be shown that the 3×1 vector of outcomes for the treatment group T before the intervention can be written as a linear combination of the 3×3 matrix M that represents the outcomes for the control groups, i.e., there exists a solution X for $M^{-1}T = X$ (see equation (16)).

$$M^{-1}T = \begin{bmatrix} 2 & 2 & -1 \\ 4 & 0 & 2 \\ 0 & 6 & -1 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 0.1 \\ 0.8 \\ 0.8 \end{bmatrix} = X \tag{16}$$

In the Synthetic Controls design, the weighting from X is applied to the control groups to create a synthetic control that matches the treatment group perfectly during the pretrend. The same weighting is then applied to the control groups after the intervention, creating a counterfactual against which the observed treatment values can be compared (see Table 12).

In this stylised example, it is possible to construct a Synthetic Control that perfectly

Table 12: Illustration of Synthetic Controls (Complete)

Time	C1	C2	C3	Synthetic Control	Treatment	Difference
0	2	2	-1	$0.1 * (2) + 0.8 * (2) + 0.8 * (-1) = 1$	1	0
1	4	0	2	$0.1 * (4) + 0.8 * (0) + 0.8 * (2) = 2$	2	0
2	0	6	-1	$0.1 * (0) + 0.8 * (6) + 0.8 * (-1) = 4$	4	0
Intervention						
3	3	-2	0	$0.1 * (3) + 0.8 * (-2) + 0.8 * (0) = -1.3$	6	7.3
4	1	4	2	$0.1 * (1) + 0.8 * (4) + 0.8 * (2) = 4.9$	7	2.1

NOTE: This table was created by the author.

matches the treatment in each of the time periods before the intervention. In practice, Synthetic Controls are typically applied when there are far more than 3 potential control groups and more than 3 time periods before the treatment takes place, and the Synthetic Control need not perfectly match the treatment group's pretrends.

When there are a large number of potential controls, the Synthetic Controls method also presents an intuitive way to test whether the treatment had an effect: placebo trials (Abadie, Diamond and Hainmueller, 2010). It is assumed that all the control groups did not have any effect from the intervention. Consequently, if the Synthetic Control method is applied to any of the control groups, any post-trend variation from the Synthetic Control exemplifies variation that may have occurred in the absense of the intervention. The researcher can apply synthetic controls to each of the control groups, capture the post-intervention variation from each of these placebo trials, and compare them against the post-intervention variation observed in the treatment group. The percentile of the treatment group compared with the controls can be considered a p-value for the effect of the intervention. For example, if there are 100 control groups, each with their own placebo trial, and the treatment group has higher variation from the synthetic control than 99 of the placebo trials, the p-value for the effect of the treatment would be $p = 0.01$.

While the Synthetic Control design allows researchers to implement an ITSA in settings with multiple control groups but no control groups with parallel pre-trends, it has several limitations. The parallel trends assumption between the synthetic control and the treatment group is the same as that between the treatment and control group in ITSA. If there are other relevant interventions occurring near the time of the tested treatment, it is challenging to disentangle the effects. The counterfactual constructed in the synthetic controls method is fictional, and so in using synthetic controls, the researcher assumes that it is reasonable to take the weighted average of different control groups.

3.3.7 Regression Discontinuity Design

Regression modelling, multi-level models, difference-in-difference, ITSA, and synthetic controls are common methods that allow researchers to make necessary adjustments for some level of confounding under the assumption that, after adjusting for these confounders, the control and treatment groups only vary in the effect of the treatment and a random error term. In certain circumstances, however, it is possible to find two samples that are so similar that it can reasonably be assumed that the distribution of all confounders (observed and unobserved) are equal between the control group. These circumstances commonly use a causal inference method called "Regression Discontinuity Design" (RDD) (Moscoe, Bor and Bärnighausen, 2015).

RDD is conceptually quite similar to ITSA, which uses the timing of an intervention as a breakpoint on the continuous variable time between the pre-trend and the post-trend. RDD is also used when there is some hard breakpoint on a continuous variable (called the "forcing variable") in the data. Importantly, the placement of the breakpoint is assumed to be entirely exogenous, i.e., not be linked with any possible confounder, and as such, it is assumed that whether an observation lies on either side of the breakpoint is entirely random (Hahn, Todd and Klaauw, 2001).

As an example, consider babies with low birth weight (LBW) in the United States (Almond et al., 2010). LBW is assessed by a certain threshold (say, 1500g) on baby weight, a continuous variable with at least some random variation. If a baby is any weight below the threshold, even 1499.9g, she is considered to have LBW; if a baby is any weight above the threshold, even 1500.1g, she does not have LBW. Babies with LBW receive additional medical services (for simplicity, say that all babies with LBW receive a procedure and babies without LBW do not). A researcher hopes to assess the impact of this procedure on the baby's weight at age 1. A classic Difference-in-Difference approach would compare the weight of all babies with and without LBW accounting for several observable confounders. However, the hard cutoff for the continuous variable presents an opportunity to adjust for observed and unobserved confounding. Those babies born within a gram on either side of the threshold are likely to be very similar in all observed and unobserved dimensions; their placement on either side of the threshold is effectively randomly assigned. Consequently, like with a randomised experiment, it may be appropriate to simply compare the outcomes of babies born on either side of the threshold.

The underlying assumption of RDD is that individuals on either side of the threshold are effectively randomly assigned. For this reason, RDD designs typically restrict their analysis to those observations that are placed very close to the threshold. In the context of the LBW example, it is intuitively easier to assume that babies born at 1499.5g are comparable to those born at 1500.5g than to assume that babies born at 700g are comparable to those born at 2300g. That is, the assumption of random assignment becomes weaker as the bandwidth from the threshold shrinks.

If a large number of individuals are truly randomised into two groups, it is theoretically not necessary to account for any confounding. While this is theoretically also possible in RDD, most RDDs also adjust for confounders. An RDD can be modelled with regression in equation (17).

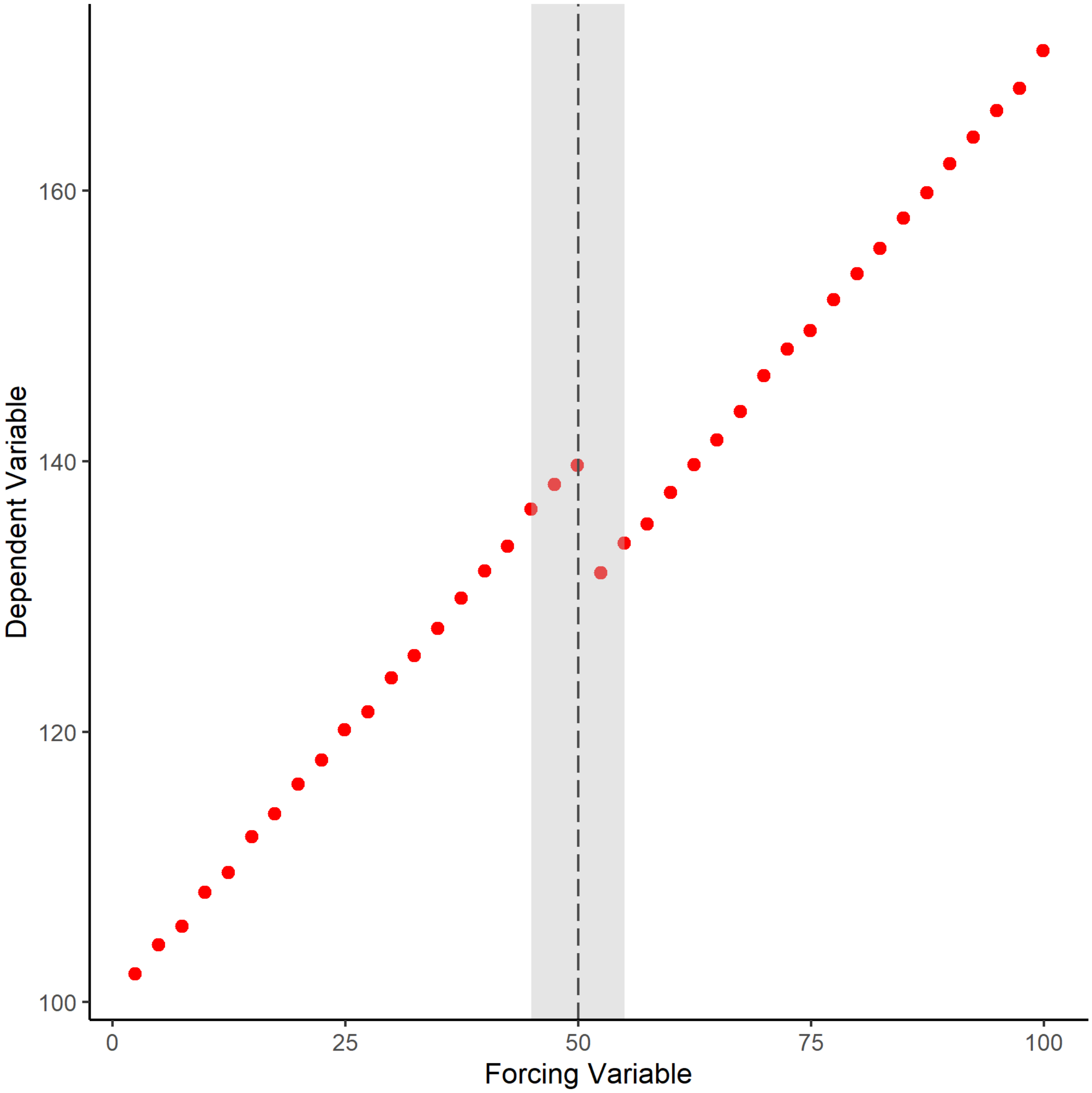
$$y_{is} = \beta_0 + \beta_1 * \text{Treatment} + X_i\beta + \epsilon_{is} \tag{17}$$

In this model, y_{is} is the outcome for individual i in condition s, Treatment refers to an indicator variable equal to 1 or 0 depending upon whether the individual was above or below the threshold, and X_i is a $N \times p$ matrix constructed by binding the columns of p observed confounders. For simplicity, assume that Treatment is equal to 1 if it is in the treatment group (e.g., birth weight lower than the LBW threshold) and 0 if it is in the control group (e.g., birth weight higher than the LBW threshold). β_0 is the parameter for the intercept, β_1 is the effect of the treatment, β is a $p \times 1$ column vector for the parameter estimates for the p observed confounders, and ϵ_{is} is a random error term. β_1 is the parameter of interest, representing the effect of the treatment.

The RDD design is shown graphically in Figure 6. Note that only those in the shaded region are included in the analysis, which compares those observations on the left hand side of the threshold within the shaded region with those on the right hand side fo the threshold within the shaded region.

RDD is a strongly internally valid method for causal inference. However, it has several limitations. First, RDD can only be applied in contexts where there is an exogenous and hard cutoff over a continuous variable, while the other causal inference methods can be applied in more general situations. Second, there is a substantial trade-off concerning the researcher's choice of bandwidth around the threshold (Bor et al., 2014). A study has increasing internal validity with a smaller bandwidth; as control and treatment groups are constructed closer together, the assumption that individual observations are randomly assigned to each becomes weaker. For example, as seen in Figure 6, it is clear that the broader the bandwidth the more biased each group becomes as they capture some of the effect of the forcing variable. On the other hand, reducing the bandwidth reduces the analytical sample size, making causal estimates less precise. If sample size constraints require a broad bandwidth, the researcher can show that the two groups are similar on observable confounders, but it becomes more difficult to rule out omitted variable bias. Third, RDD methods typically lack external validiity (Venkataramani, Bor and Jena, 2016). Because only those observations right above and below the threshold are considered in the analysis, the results cannot be reliably extrapolated to observations falling further away from the threshold. For example,

Figure 6: Illustration of Regression Discontinuity Design



NOTE: This figure was generated from simulations by the author. The red dots represent groups of observations. The threshold, represented by the dashed black line, is set at 50, and the bandwidth, represented by the gray rectangular area, is set with a radius of 5. The initial intercept is set at 100 and the initial slope is set at 0.8. The change in level at the intervention (i.e., the true effect) is set at −10. There is no change in slope after the intervention (i.e., the effect of the forcing variable, aside from the threshold, is constant).

in the example of babies with LBW, the effect of the procedure on babies who are much lower than the threshold cannot be reliably discerned. The effect estimated RDD studies is, consequently, often described as the Local Average Treatment Effect (LATE) rather than the Average Treatment Effect (**?**).

3.3.8 Instrumental Variable

RDD exploits some random variation – an exogenous threshold over the forcing variable – to mitigate the threat of confounding and draw causal inference. Similarly, the Instrumental Variable method uses an exogenous source of variation to control for confounding (Angrist and Krueger, 2001).

The Instrumental Variable approach relies upon finding a third variable, called the "instrument", other than the outcome or the exposure that (A) is correlated with the exposure, (B) is only correlated with the outcome through the exposure, and (C) is otherwise exogenous from the system (Greenland, 2000). Conceptually, because the instrument is exogenous to all potential confounders and only affects the outcome through the exposure, it is possible to discern the causal effect of the exposure through the relationship between the instrument and the outcome. Say, for example, a 1-unit increase in the instrument is related to a 0.6-unit increase in the exposure and a 0.3-unit increase in the outcome. Assuming that the instrument is only correlated with the outcome through the exposure and unrelated to all confounders, then a 0.6-unit increase in the exposure causally increases the outcome by 0.3-units (an effect size of 0.5).

Instrumental variables are often modelled using systems of equations that are each computed via regression, such as those shown in equations (18) and (19).

$$x = \pi_0 + \pi_1 * z + u \tag{18}$$

$$y = \beta_0 + \beta_1 * \hat{x} + X\beta + v \tag{19}$$

In equation (18), x represents the exposure variable and z represents the instrument variable. In equation (19), y is the outcome, $\hat{x}$ is the values of the exposure predicted from equation (18), X is a $N \times p$ matrix constructed by binding p columns of confounders together. The parameter of interest is β_1 in equation (19), which corresponds to the causal effect of the exposure on the outcome.

The IV approach is, overall, a strong method for causal inference, but it has several limitations. Perhaps most importantly, the instrumental variable approach only works in settings where an appropriate instrument exists and can be observed. Further, the researcher needs to make some assumptions to assert that a given variable is, in fact, an appropriate instrument. It is possible for the researcher to empirically test (A) whether the instrument is correlated with the exposure, (B) whether the instrument is uncorrelated with the observed confounders, and (C) whether the instrument is only correlated with the outcome through the exposure (e.g., there is no relationship between the instrument and the outcome after adjusting for the exposure) (Zhang et al., 2018). However, it remains possible that the instrument is still endogenous (e.g., correlated with an unobserved confounder) (Bascle, 2008). In this case, the IV estimator may be biased. Further, the IV approach is less efficient than single-step methods like OLS, leading to larger standard errors for the parameter estimates (Semadeni, Withers and Certo, 2014).

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. “Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program.” *Journal of the American Statistical Association*, 105(490): 493–505.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2015. “Comparative Politics and the Synthetic Control Method.” *American Journal of Political Science*, 59(2): 495–510.

Addicott, Rachael, and Christopher Ham. 2014. *Commissioning and funding general practice: making the case for family care networks.* OCLC: 871306464.

Almond, Douglas, Joseph J. Doyle, Amanda E. Kowalski, and Heidi Williams. 2010. “Estimating Marginal Returns to Medical Care: Evidence from At-Risk Newborns.” *The quarterly journal of economics*, 125(2): 591–634.

Angrist, Joshua D., and Alan B. Krueger. 2001. “Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments.” *Journal of Economic Perspectives*, 15(4): 69–85.

Anonymous. 1953. “General Practice Today and Tomorrow.” *British Medical Journal*, 717–719.

Anthony, E. 1950. “The G.P. at the Crossroads.” *British Medical Journal*, 1(4661): 1077–1079.

Applebee, Kathie. 2006. “Understanding the revised QOF.” *Independent Nurse*, 2006(5).

Arnold, S. R., and S. E. Straus. 2006. “Interventions to improve antibiotic prescribing practices in ambulatory care.” *Evidence-Based Child Health: A Cochrane Review Journal*, 1(2): 623–690.

Aveling, Emma-Louise, Graham Martin, Natalie Armstrong, Jay Banerjee, and Mary Dixon-Woods. 2012. "Quality improvement through clinical communities: eight lessons for practice." *Journal of Health Organization and Management*, 26(2): 158–174.

Avorn, Jerry, and Stephen B. Soumerai. 1983. "Improving Drug-Therapy Decisions through Educational Outreach." *New England Journal of Medicine*, 308(24): 1457–1463.

Baker, R. H. 1989. "The quality initiative of the Royal College of General Practitioners." *Quality Assurance in Health Care: The Official Journal of the International Society for Quality Assurance in Health Care*, 1(1): 23–29.

Baker, R., M. Lakhani, R. Fraser, and F. Cheater. 1999. "A model for clinical governance in primary care groups." *BMJ*, 318(7186): 779–783.

Bandura, Albert. 1986. *Social foundations of thought and action: A social cognitive theory. Social foundations of thought and action: A social cognitive theory*, Englewood Cliffs, NJ, US:Prentice-Hall, Inc.

Bascle, Guilhem. 2008. "Controlling for endogeneity with instrumental variables in strategic management research." *Strategic Organization*, 6(3): 285–327.

Beardow, R., K. Cheung, and W. M. Styles. 1993. "Factors influencing the career choices of general practitioner trainees in North West Thames Regional Health Authority." *British Journal of General Practice*, 43(376): 449–452.

Bell, Andrew, Malcolm Fairbrother, and Kelvyn Jones. 2019. "Fixed and random effects models: making an informed choice." *Quality & Quantity*, 53(2): 1051–1074.

Bennett, Brian, Emilene Coventry, Nicola Greenway, and Mark Minchin. 2014. "The NICE process for developing quality standards and indicators." *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, 108(8): 481–486.

Berwick, Donald, and Daniel M. Fox. 2016. "“Evaluating the Quality of Medical Care”: Donabedian's Classic Article 50 Years Later." *The Milbank Quarterly*, 94(2): 237–241.

Bevan, Gwyn. 2010. "Performance Measurement of "Knights" and "Knaves": Differences in Approaches and Impacts in British Countries after Devolution." *Journal of Comparative Policy Analysis: Research and Practice*, 12(1-2): 33–56.

Bevan, Gwyn, and Barbara Fasolo. 2013. "Models of governance of public services: empirical and behavioural analysis of 'econs' and 'humans'." In *Behavioural Public Policy..* , ed. Oliver Angus, 38–62. Cambridge, UK:Cambridge University Press.

Bird, Sheila M., Cox Sir David, Vern T. Farewell, Goldstein Harvey, Holt Tim, and Smith Peter C. 2005. "Performance indicators: good, bad, and ugly." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(1): 1–27.

Blackwell, Matthew, Stefano Iacus, Gary King, and Giuseppe Porro. 2009. "Cem: Coarsened Exact Matching in Stata." *The Stata Journal*, 9(4): 524–546.

Bor, Jacob, Ellen Moscoe, Portia Mutevedzi, Marie-Louise Newell, and Till Bärnighausen. 2014. "Regression Discontinuity Designs in Epidemiology." *Epidemiology (Cambridge, Mass.)*, 25(5): 729–737.

Boulkedid, Rym, Hendy Abdoul, Marine Loustau, Olivier Sibony, and Corinne Alberti. 2011. "Using and Reporting the Delphi Method for Selecting Healthcare Quality Indicators: A Systematic Review." *PLOS ONE*, 6(6): e20476.

British Medical Journal. 1985. "“Towards quality in general practice”: GMSC discusses RCGP's paper." *Br Med J (Clin Res Ed)*, 291(6500): 987–988.

Campbell, S M. 2002. "Research methods used in developing and applying quality indicators in primary care." *Quality and Safety in Health Care*, 11(4): 358–364.

Campbell, Stephen, and Helen Lester. 2010. "Developing indicators and the concept of QOFability." In *The Quality and Outcomes Framework: QOF - Transforming General Practice.* , ed. Stephen Gillam and A. Niroshan Siriwardena, 16–27. Radcliffe Publishing.

Campbell, Stephen M., Evangelos Kontopantelis, Kerin Hannon, Martyn Burke, Annette Barber, and Helen E. Lester. 2011. "Framework and indicator testing protocol for developing and piloting quality indicators for the UK quality and outcomes framework." *BMC Family Practice*, 12(1): 85.

Carlisle, R., and S. Johnstone. 1996. "Factors influencing the response to advertisements for general practice vacancies." *BMJ*, 313(7055): 468–471.

Cheng, Lim Puay, Nelson K.H. Tang, and Peter M. Jackson. 1999. "An innovative framework for health care performance measurement." *Managing Service Quality: An International Journal*, 9(6): 423–433.

Chien, Alyna T., and Meredith B. Rosenthal. 2013. "Medicare's Physician Value-Based Payment Modifier — Will the Tectonic Shift Create Waves?" *New England Journal of Medicine*, 369(22): 2076–2078.

Chung, Paul J, Jeanette Chung, Manish N. Shah, and David O. Meltzer. 2003. "How Do Residents Learn? The Development of Practice Styles in a Residency Program." *Ambulatory Pediatrics*, 3(4): 166–172.

Cockburn, J., D. Ruth, C. Silagy, M. Dobbin, Y. Reid, M. Scollo, and L. Naccarella. 1992. "Randomised trial of three approaches for marketing smoking cessation programmes to Australian general practitioners." *British Medical Journal*, 304(6828): 691–694.

Collings, Joseph S. 1950. "General Practice in England Today - A Reconnaisance." *The Lancet*, 255(6604): 555.

Conn, David. 1982. "Effort, efficiency, and incentives in economic organizations." *Journal of Comparative Economics*, 6(3): 223–234.

Conry, Mary C, Niamh Humphries, Karen Morgan, Yvonne McGowan, Anthony Montgomery, Kavita Vedhara, Efharis Panagopoulou, and Hannah Mc Gee. 2012. "A 10year (2000–2010) systematic review of interventions to improve quality of care in hospitals." *BMC Health Services Research*, 12: 275.

Craig, Peter. 2015. "Synthetic Controls: A New Approach to Evaluating Interventions."

Craig, Peter, Cyrus Cooper, David Gunnell, Sally Haw, Kenny Lawson, Sally Macintyre, David Ogilvie, Mark Petticrew, Barney Reeves, Matt Sutton, and Simon Thompson. 2012. "Using natural experiments to evaluate population health interventions: new Medical Research Council guidance." *J Epidemiol Community Health*, 66(12): 1182–1186.

Crémieux, Pierre-Yves, and Pierre Ouellette. 2001. "Omitted variable bias and hospital costs." *Journal of Health Economics*, 20(2): 271–282.

Dambrun, Michaël, Serge Guimond, and Sandra Duarte. 2002. "The impact of hierarchy-enhancing vs. attenuating academic major on stereotyping: The mediating role of perceived social norm." *Current Research in Social Psychology*, 7(8): 114–136.

Dawson, G. de H. 1950. "The G.P. at the Crossroads." *British Medical Journal*, 1(4667): 1431.

de Jong, Judith D, Peter P Groenewegen, and Gert P Westert. 2003. "Mutual influences of general practitioners in partnerships." *Social Science & Medicine*, 57(8): 1515–1524.

Deyo-Svendsen, Mark E., Michael R. Phillips, Jill K. Albright, Keith A. Schilling, and Karl B. Palmer. 2016. "A Systematic Approach to Clinical Peer Review in a Critical Access Hospital." *Quality Management in Healthcare*, 25(4): 213.

Dixon, Anna, Artak Khachatryan, Andrew Wallace, Stephen Peckham, Tammy Boyce, and Stephen Gillam. 2011. "The Quality and Outcomes Framework (QOF): does it reduce health inequalities?"

Donabedian, Avedis. 1966. "Evaluating the Quality of Medical Care." *The Milbank Memorial Fund Quarterly*, 44(3): 166–206.

Donabedian, Avedis. 1980. *Explorations in Quality Assessment and Monitoring: The Definition of Quality and Approaches to Its Assessment: 1.* . Spi edition ed., Ann Arbor, Mich:Health Administration Press.

Doran, Tim, Evangelos Kontopantelis, David Reeves, Matthew Sutton, and Andrew M. Ryan. 2014. "Setting performance targets in pay for performance programmes: what can we learn from QOF?" *BMJ*, 348.

Downing, Amy, Gavin Rudge, Yaping Cheng, Yu-Kang Tu, Justin Keen, and Mark S. Gilthorpe. 2007. "Do the UK government's new Quality and Outcomes Framework (QOF) scores adequately measure primary care performance? A cross-sectional survey of routine healthcare data." *BMC Health Services Research*, 7(1): 166.

Duncan, Craig, Kelvyn Jones, and Graham Moon. 1996. "Health-related behaviour in context: A multilevel modelling approach." *Social Science & Medicine*, 42(6): 817–830.

Eddy, David M. 1984. "Variations in Physician Practice: The Role of Uncertainty." *Health Affairs*, 3(2): 74–89.

Eisenberg, J. M. 1985. "Physician utilization. The state of research about physicians' practice patterns." *Medical Care*, 23(5): 461–483.

Endsley, Scott, Margaret Kirkegaard, and Anthony Linares. 2005. "Working Together: Communities of Practice in Family Medicine."

Erickson, Bonnie H. 1988. "The relational basis of attitudes." In *Social Structures: A Network Approach.* , ed. Barry Wellman, S. D. Berkowitz and Mark Granovetter. CUP Archive.

Evans, Elizabeth, Harry Aiking, and Adrian Edwards. 2011. "Reducing variation in general practitioner referral rates through clinical engagement and peer review of referrals: a service improvement project." *Quality in Primary Care*, 19(4): 263–272.

Fehr, Beverley. 1996. *Friendship Processes.* SAGE.

Fisher, R. A. 1919. "The causes of human variability." *The Eugenics Review*, 10(4): 213–220.

Fisher, R.A. 1925. *Statistical Methods for Research Workers.* London:Oliver & Boyd.

Fotaki, Marianna. 2013. "Is Patient Choice the Future of Health Care Systems?" *International Journal of Health Policy and Management*, 1(2): 121–123.

Freeman, Tim. 2002. "Using performance indicators to improve health care quality in the public sector: a review of the literature." *Health Services Management Research*, 15(2): 126–137.

Fry, John. 1988. "General Practice and Primary Health Care 1940s-1980s."

Geneau, Robert, Pascale Lehoux, Raynald Pineault, and Paul Lamarche. 2008. "Understanding the work of general practitioners: a social science perspective on the context of medical decision making in primary care." *BMC Family Practice*, 9(1): 12.

Gillam, Stephen. 2017. "The Family Doctor Charter: 50 years on." *British Journal of General Practice*, 67(658): 227–228.

Gillam, Stephen, and Aloysius Niroshan Siriwardena, ed. 2011. *Quality and outcomes framework: QOF - transforming general practice.* Abingdon:Radcliffe Pub. OCLC: ocn680639366.

Gillam, Stephen, and Nicholas Steel. 2013. "The Quality and Outcomes Framework—where next?" *BMJ*, 346.

Glover, J. Alison. 1938. "The Incidence of Tonsillectomy in School Children." *Proceedings of the Royal Society of Medicine*, 31(10): 1219–1236.

GMS Committee. 1965. "Charter for General Practice." *Br Med J*, 1(5436): 669–670.

Goldberg, B. A. 1984. "The peer review privilege: a law in search of a valid policy." *American Journal of Law & Medicine*, 10(2): 151–167.

Goodwin, James S., Yu-Li Lin, Siddhartha Singh, and Yong-Fang Kuo. 2013. "Variation in length of stay and outcomes among hospitalized patients attributable to hospitals and hospitalists." *Journal of General Internal Medicine*, 28(3): 370–376.

Goodwin, Nick, Anna Dixon, Teresa Poole, and Veena Raleigh. 2011. *Improving the quality of care in general practice: report of an independent inquiry commissioned by the King's Fund.* London:King's Fund. OCLC: 753627495.

Gosden, Toby, Isobel Bowler, and Matthew Sutton. 2000. "How Do General Practitioners Choose Their Practice? Preferences for Practice and Job Characteristics." *Journal of Health Services Research & Policy*, 5(4): 208–213.

Greenland, Sander. 2000. "An introduction to instrumental variables for epidemiologists." *International Journal of Epidemiology*, 29(4): 722–729.

Griffiths, M, W E Waters, and E D Acheson. 1979. "Variation in hospital stay after inguinal herniorrhaphy." *British Medical Journal*, 1(6166): 787–789.

Grimes, R M, and S K Moseley. 1976. “An approach to an index of hospital performance.” *Health Services Research*, 11(3): 288–301.

Guthrie, Bruce, Gary McLean, and Matt Sutton. 2006. “Workload and reward in the Quality and Outcomes Framework of the 2004 general practice contract.” *British Journal of General Practice*, 56(532): 836–841.

Hadfield, Stephen J. 1953. “A Field Survey of General Practice, 1951-2.” *British Medical Journal*, 2(4838): 683–706.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. “Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design.” *Econometrica*, 69(1): 201–209.

Holman, William L., Richard M. Allman, Monique Sansom, Catarina I. Kiefe, Eric D. Peterson, Kevin J. Anstrom, Steadman S. Sankey, Steve G. Hubbard, Robert G. Sherrill, and for the Alabama CABG Study Group. 2001. “Alabama Coronary Artery Bypass Grafting Project: Results of a Statewide Quality Improvement Initiative.” *JAMA*, 285(23): 3003.

Honigsbaum, F. 1979. “The division in British medicine : A history of the separation of general practice from hospital care, 1911-1968.” PhD diss. London School of Economics and Political Science (University of London).

Horder, J. P., and G. Swift. 1979. “The history of vocational training for general practice.” *The Journal of the Royal College of General Practitioners*, 29(198): 24–32.

Howard, David H. 2006. “Quality and Consumer Choice in Healthcare: Evidence from Kidney Transplantation.” *The B.E. Journal of Economic Analysis & Policy*, 5(1).

Howie, J. G. R., D. Heaney, and M. Maxwell. 2004. “Quality, core values and the general practice consultation: issues of definition, measurement and delivery.” *Family Practice*, 21(4): 458–468.

Hulscher, M E, M Wensing, R P Grol, T van der Weijden, and C van Weel. 1999. “Interventions to improve the delivery of preventive services in primary care.” *American Journal of Public Health*, 89(5): 737–746.

Hunt, J. H. 1955. “The Scope and Development of General Practice in Relation to Other Branches of Medicine: A Constructive Review.” *The Lancet*, 266(6892): 681–687.

Iacus, Stefano M., Gary King, and Giuseppe Porro. 2009. “cem: Software for Coarsened Exact Matching.” *Journal of Statistical Software*, 30(9).

Iedema, Rick, Shannon Meyerkort, and Les White. 2005. “Emergent modes of work and communities of practice.” *Health Services Management Research*, 18(1): 13–24.

Imai, Kosuke, Gary King, and Elizabeth A. Stuart. 2008. “Misunderstandings between experimentalists and observationalists about causal inference.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(2): 481–502.

Institute of Healthcare Management. 2018. “Focus on GP quality indicators: General Practitioners Committee.”

Institute of Medicine. 2006. *Performance Measurement: Accelerating Improvement (Pathways to Quality Health Care Series).* Washington, D.C.:National Academies Press.

Inui, Thomas S. 1976. “Improved Outcomes in Hypertension After Physician Tutorials: A Controlled Trial.” *Annals of Internal Medicine*, 84(6): 646.

Jaffer, Amir K., Daniel J. Brotman, Lori D. Bash, Syed K. Mahmood, Brooke Lott, and Richard H. White. 2010. “Variations in perioperative warfarin management:

outcomes and practice patterns at nine hospitals." *The American Journal of Medicine*, 123(2): 141–150.

Jamtvedt, G., J. M. Young, D. T. Kristoffersen, M. A. Thomson O'Brien, and A. D. Oxman. 2003. "Audit and feedback: effects on professional practice and health care outcomes." *The Cochrane Database of Systematic Reviews*, , (3): CD000259.

Jarvis, William R. 2012. "What can Canada learn from the USA's experience in reducing healthcare-associated infections?" *Clinical Governance: An International Journal*, 17(2): 149–154.

Jensen, Michael C., and William H. Meckling. 1976. "Theory of the firm: Managerial behavior, agency costs and ownership structure." *Journal of Financial Economics*, 3(4): 305–360.

Johnson, Bradford C., James M. Manyika, and Lareina A. Yee. 2005. "The next revolution in interactions." McKinsey Quarterly.

Jong, Judith D. de, Gert P. Westert, Ronald Lagoe, and Peter P. Groenewegen. 2006. "Variation in Hospital Length of Stay: Do Physicians Adapt Their Length of Stay Decisions to What Is Usual in the Hospital Where They Work?" *Health Services Research*, 41(2): 374–394.

Jordan, Michelle E., Holly J. Lanham, Benjamin F. Crabtree, Paul A. Nutting, William L. Miller, Kurt C. Stange, and Reuben R. McDaniel. 2009. "The role of conversation in health care interventions: enabling sensemaking and learning." *Implementation Science*, 4(1): 15.

Joss, Richard, and Maurice Kogan. 1995. *Advancing Quality: Total Quality Management in the NHS.* Buckingham ; Bristol, PA, USA:Open University Press.

Keating, Nancy L., John Z. Ayanian, Paul D. Cleary, and Peter V. Marsden. 2007. “Factors affecting influential discussions among physicians: a social network analysis of a primary care practice.” *Journal of General Internal Medicine*, 22(6): 794–798.

Keating, N. L., A. M. Zaslavsky, and J. Z. Ayanian. 1998. “Physicians’ experiences and beliefs regarding informal consultation.” *JAMA*, 280(10): 900–904.

Krasnik, A., P. P. Groenewegen, P. A. Pedersen, P. von Scholten, G. Mooney, A. Gottschau, H. A. Flierman, and M. T. Damsgaard. 1990. “Changing remuneration systems: effects on activity in general practice.” *British Medical Journal*, 300(6741): 1698–1701.

Lask, M. 1950. “The G.P. at the Crossroads.” *British Medical Journal*, 1(4658): 902.

Lechner, Michael. 2010. “The Estimation of Causal Effects by Difference-in-Difference Methods.” *Foundations and Trends® in Econometrics*, 4(3): 165–224.

Le Grand, Julian. 2007. *The other invisible hand: delivering public services through choice and competition.* Princeton:Princeton University Press.

Lester, Helen, and Stephen Campbell. 2010. “Developing Quality and Outcomes Framework (QOF) indicators and the concept of ’QOFability’.” *Quality in Primary Care*, 18(2): 103–109.

Lester, Helen, Deborah J. Sharp, F. D. R. Hobbs, and Mayur Lakhani. 2006. “The Quality and Outcomes Framework of the GMS contract: a quiet evolution for 2006.” *British Journal of General Practice*, 56(525): 244–246.

Leyland, Alastair H., and Peter P. Groenewegen. 2003. “Multilevel modelling and public health policy.” *Scandinavian Journal of Public Health*, 31(4): 267–274.

Lipman, Toby, Chris Johnstone, Martin Roland, Patricia Wilkie, and David Jewell. 2005. “So how was it for you? A year of the GMS ContractNever offer GPs money, they will just take itAn important step forwardsIs the GMS contract just for doctors? Or do patients benefit as well?Careful with the unintended consequences: INTO THE SUNLIT UPLANDS?" *British Journal of General Practice*, 55(514): 396–396.

Lohr, Kathleen N., ed. 1990. *Medicare: A Strategy for Quality Assurance: Volume 1.* Washington (DC):National Academies Press (US).

Lopez Bernal, James, Steven Cummins, and Antonio Gasparrini. 2016. "Interrupted time series regression for the evaluation of public health interventions: a tutorial." *International Journal of Epidemiology*, dyw098.

Lopez Bernal, James, Steven Cummins, and Antonio Gasparrini. 2018. "The use of controls in interrupted time series studies of public health interventions." *International Journal of Epidemiology*, 47(6): 2082–2093.

Magee, Helen, Lucy-Jane Davis, and Angela Coulter. 2003. "Public Views on Healthcare Performance Indicators and Patient Choice." *Journal of the Royal Society of Medicine*, 96(7): 338–342.

Mainz, J. 2003*a*. "Defining and classifying clinical indicators for quality improvement." *International Journal for Quality in Health Care*, 15(6): 523–530.

Mainz, Jan. 2003*b*. "Developing evidence-based clinical indicators: a state of the art methods primer." *International Journal for Quality in Health Care: Journal of the International Society for Quality in Health Care*, 15 Suppl 1: i5–11.

Meltzer, David, Jeanette Chung, Parham Khalili, Elizabeth Marlow, Vineet Arora, Glen Schumock, and Ron Burt. 2010. "Exploring the use of social network methods in designing healthcare quality improvement teams." *Social Science & Medicine*, 71(6): 1119–1130.

Mittman, Brian S., Xenia Tonesk, and Peter D. Jacobson. 1992. “Implementing Clinical Practice Guidelines: Social Influence Strategies and Practitioner Behavior Change.” *QRB - Quality Review Bulletin*, 18(12): 413–422.

Mohammad, Mosadeghrad Ali. 2013. “Healthcare service quality: towards a broad definition.” *International Journal of Health Care Quality Assurance*, 26(3): 203–219.

Mosadeghrad, Ali Mohammad. 2012. “A Conceptual Framework for Quality of Care.” *Materia Socio-Medica*, 24(4): 251–261.

Mosadeghrad, Ali Mohammad. 2014*a*. “Factors Affecting Medical Service Quality.” *Iranian Journal of Public Health*, 43(2): 210–220.

Mosadeghrad, Ali Mohammad. 2014*b*. “Factors influencing healthcare service quality.” *International Journal of Health Policy and Management*, 3(2): 77–89.

Moscoe, Ellen, Jacob Bor, and Till Bärnighausen. 2015. “Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice.” *Journal of Clinical Epidemiology*, 68(2): 132–143.

National Health Service. 2005. “National Quality and Outcomes Framework Statistics for England - 2004-05.”

National Health Service. 2006*a*. “National Quality and Outcomes Framework Achievement Data - England, 2005-06.”

National Health Service. 2006*b*. “NHS Employers: Directed Enhanced Services.”

National Health Service. 2010. “Quality and Outcomes Framework - 2009-10.”

National Health Service. 2014. “Quality and Outcomes Framework (QOF) - 2013-14 - NHS Digital.”

National Health Service. 2018*a*. “QOF 2018/19.”

National Health Service. 2018*b*. "Quality and Outcomes Framework, Achievement, prevalence and exceptions data - 2017-18 [PAS]."

National Health Service. 2019. "Quality and Outcomes Framework, Achievement, prevalence and exceptions data 2018-19 [PAS] - NHS Digital."

National Institute for Health and Care Excellence. 2017. "Health and Social Care Directorate Indicator Process Guide."

NHS England. 2019. "2019/20 General Medical Services (GMS) contract Quality and Outcomes Framework (QOF)."

NICE. 2019. "Quality and Outcomes Framework Indicators."

O'Connor, Gerald T., Stephen K. Plume, Elaine M. Olmstead, Jeremy R. Morton, Christopher T. Maloney, William C. Nugent, Felix Hernandez, Robert Clough, Bruce J. Leavitt, Laurence H. Coffin, Charles A. S. Marrin, David Wennberg, John D. Birkmeyer, David C. Charlesworth, David J. Malenka, Hebe B. Quinton, Joseph F. Kasper, Felix Hernandez, Hebe Quinton, Deborah Carey-Johnson, Cynthia M. Downs, Joseph J. Hessel, Robert M. Hoffman, Edward R. Johnson, Helen McKinnon, Cathy Mingo, Craig Pedersen, Wendy Perkins, Matthew L. Rowe, Katrina Sargent, Ted Silver, Peter Ver Lee, Craig Warren, Shelley Barber, Pamela Brown, Richard G. Brandenburg, Steve Colmanaro, Walter D. Gundel, Richard S. Jackson, David Johnson, Ann Laramee, William C. Paganelli, Diane Pappalardo, Daniel S. Raabe, Christopher Terrien, Paul Vaitkus, Matthew Watkins, Virginia Beggs, William Burke, Edward Catherwood, Lawrence J. Daeey, Candis Darcey, Gordon Defoe, Thomas Dodds, Mary Fillinger, Bruce Friedman, Bruce Hettleman, Terry Kneeland, Elizabeth Maislen, Nathaniel Niles, Nancy J. O'Connor, John Robb, William Schults, William F. Sullivan, Jon Wahrenberger, Beth Wolf, Yvon

Baribeau, Ann Becker, Craig C. Berry, Kevin Berry, William A. Bradley, S. Cuddy, Robert C. Dewey, Frank Fedele, Louis I. Fink, Erik J. Funk, Alan E. Garstka, Dan Halstead, Michael J. Hearne, J. Beatty Hunter, Alan D. Kaplan, Peggy Lambert, Patrick M. Lawrence, Jeffery Lockhart, Kathy McNeil, Edward Palank, M. Judith Porelle, Donna Pulsifer, Joanne Robichaud, James Schnitz, Shirley Shea, Benjamin M. Westbrook, Thomas P. Wharton, Kirke W. Wheeler, Dee White, Robert B. Keller, David C. Soule, Mary Abbott, Lawrence Adrian, Warren D. Alpern, Eric Anderson, Richard A. Anderson, Linda Banister, Claire Berg, Seth Blank, Carl E. Bredenberg, Michael Brennan, Linda Brewster, David Burkey, Alice Cirillo, Cantwell Clark, Jane Cleaves, Deborah Courtney, Joshua Cutler, Desmond Donegal, Pat Fallo, Daniel Hanley, Jane Kane, Saul Katz, Mirle A. Kellett, Robert Kramer, Costas T. Lambrew, F. Stephen Larned, Chris A. Lutes, Edward R. Nowicki, John R. O'Meara, Harold Osher, Patricia Peasley, Cathy Prouty, Reed D. Quinn, Dennis Redfield, Karen Reynolds, Thomas Ryan, Jean Saunders, Alyce Schultz, Susan Seekins, Russell Stogsdill, Paul W. Sweeny, Karen Tolan, Nancy Tooker, Joan F. Tryzelaar, Marie Turcotte, Kathy Viger, Cynthia Westlund, Richard L. White, Wanda Whittet, and Carol Zografos. 1996. "A Regional Intervention to Improve the Hospital Mortality Associated With Coronary Artery Bypass Graft Surgery." *JAMA*, 275(11): 841–846.

O'Connor, G. T., H. B. Quinton, N. D. Traven, L. D. Ramunno, T. A. Dodds, T. A. Marciniak, and J. E. Wennberg. 1999. "Geographic variation in the treatment of acute myocardial infarction: the Cooperative Cardiovascular Project." *JAMA*, 281(7): 627–633.

Ovretveit, John, and Christina Townsend. 1992. *Health service quality: an introduction to quality methods for health services.* Oxford:Blackwell Science. OCLC: 934391516.

Ozcan, Yasar A. 2005. *Quantitative Methods in Health Care Management: Techniques and Applications.* John Wiley & Sons.

Paul-Shaheen, P., J. D. Clark, and D. Williams. 1987. "Small area analysis: a review and analysis of the North American literature." *Journal of Health Politics, Policy and Law*, 12(4): 741–809.

Pencheon, David. 2007. "The Good Indicators Guide: Understanding How to Use and Choose Indicators."

Penfold, Robert B., and Fang Zhang. 2013. "Use of Interrupted Time Series Analysis in Evaluating Health Care Quality Improvements." *Academic Pediatrics*, 13(6, Supplement): S38–S44.

Pescosolido, Bernice A. 1992. "Beyond Rational Choice: The Social Dynamics of How People Seek Help." *American Journal of Sociology*, 97(4): 1096–1138.

Petchey, R., J. Williams, and M. Baker. 1997. "'Ending up a GP': a qualitative study of junior doctors' perceptions of general practice as a career." *Family Practice*, 14(3): 194–198.

Petchey, Roland. 1995. "Collings report on general practice in England in 1950: unrecognised, pioneering piece of British social research?" *BMJ*, 311(6996): 40–42.

Pirrie, Ivan. 1950. "The G.P. at the Crossroads." *British Medical Journal*, 2(4671): 169–170.

Powell, J. Enoch. 1976. *Medicine and politics: 1975 and after.* . New ed ed., Tunbridge Wells:Pitman Medical.

Practice, British Journal of General. 1985. "Quality assured." *The Journal of the Royal College of General Practitioners*, 35(278): 411–411.

Proctor, and Wright. 1998. "Can services marketing concepts be applied to health care?" *Journal of Nursing Management*, 6(3): 147–153.

Pronovost, Peter J., Sam R. Watson, Christine A. Goeschel, Robert C. Hyzy, and Sean M. Berenholtz. 2016. "Sustaining Reductions in Central Line–Associated Bloodstream Infections in Michigan Intensive Care Units: A 10-Year Analysis." *American Journal of Medical Quality*, 31(3): 197–202.

Puntanen, Simo, and George P. H. Styan. 1989. "The Equality of the Ordinary Least Squares Estimator and the Best Linear Unbiased Estimator." *The American Statistician*, 43(3): 153–161.

Pérez-Cuevas, Ricardo, Hector Guiscafré, Onofre Muñoz, Hortensia Reyes, Patricia Tomé, Vita Libreros, and Gonzalo Gutiérrez. 1996. "Improving physician prescribing patterns to treat rhinopharyngitis. Intervention strategies in two health systems of Mexico." *Social Science & Medicine*, 42(8): 1185–1194.

Raleigh, Veena, and Catherine Foot. 2010. *Getting the measure of quality: opportunities and challenges.* London:King's Fund. OCLC: 502427584.

Rawlins, Michael. 1999. "In pursuit of quality: the National Institute for Clinical Excellence." *The Lancet*, 353(9158): 1079–1082.

Reid, Constance. 1998. "Neyman—from life." In *Neyman.* , ed. Constance Reid, 1–293. New York, NY:Springer.

Roland, Martin. 2007. "The Quality and Outcomes Framework: too early for a final verdict." *The British Journal of General Practice*, 57(540): 525–527.

Roland, Martin, and Bruce Guthrie. 2016. "Quality and Outcomes Framework: what have we learnt?:." *BMJ*, i4060.

Roland, Martin, and Stephen Campbell. 2014. "Successes and Failures of Pay for Performance in the United Kingdom." *New England Journal of Medicine*, 370(20): 1944–1949.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika*, 70(1): 41–55.

Rosenbaum, Paul R., and Donald B. Rubin. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association*, 79(387): 516–524.

Rosenbaum, Paul R., and Donald B. Rubin. 1985. "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score." *The American Statistician*, 39(1): 33–38.

Rowsell, R., M. Morgan, and J. Sarangi. 1995. "General practitioner registrars' views about a career in general practice." *British Journal of General Practice*, 45(400): 601–604.

Royal College of General Practitioners. 2019. "History of the College."

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*, 66(5): 688–701.

Rubin, Donald B. 2005. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions." *Journal of the American Statistical Association*, 100(469): 322–331.

Sanguinetti, Harold H. 1950. "The G.P. at the Crossroads." *British Medical Journal*, 1(4664): 1270–1271.

Sargeant, Joan, Jocelyn Lockyer, Karen Mann, Eric Holmboe, Ivan Silver, Heather Armson, Erik Driessen, Tanya MacLeod, Wendy Yen, Kathryn

Ross, and Mary Power. 2015. "Facilitated Reflective Performance Feedback: Developing an Evidence- and Theory-Based Model That Builds Relationship, Explores Reactions and Content, and Coaches for Performance Change (R2C2)." *Academic Medicine*, 90(12): 1698–1706.

Schuster, Mark A., Elizabeth A. McGlynn, and Robert H. Brook. 1998. "How Good Is the Quality of Health Care in the United States?" *The Milbank Quarterly*, 76(4): 517–563.

Scott, Anthony. 2001. "Eliciting GPs' preferences for pecuniary and non-pecuniary job characteristics." *Journal of Health Economics*, 20(3): 329–347.

Scott, Anthony, Julia Witt, John Humphreys, Catherine Joyce, Guyonne Kalb, Sung-Hee Jeon, and Matthew McGrail. 2013. "Getting doctors into the bush: General Practitioners' preferences for rural location." *Social Science & Medicine*, 96: 33–44.

Secretary of State for Health. 1998. "A First Class Service."

Semadeni, Matthew, Michael C. Withers, and S. Trevis Certo. 2014. "The perils of endogeneity and instrumental variables in strategy research: Understanding through simulations." *Strategic Management Journal*, 35(7): 1070–1079.

Shekelle, Paul. 2003. "New contract for general practitioners: A bold initiative to improve quality of care, but implementation will be difficult." *BMJ*, 326(7387): 457–458.

Shekelle, Paul G. 2013. "Quality indicators and performance measures: methods for development need more standardization." *Journal of Clinical Epidemiology*, 66(12): 1338–1339.

Starkweather, D. B., L. Gelwicks, and R. Newcomer. 1975. "Delphi forecasting of health care organization." *Inquiry: A Journal of Medical Care Organization, Provision and Financing*, 12(1): 37–46.

Stelfox, Henry T., and Sharon E. Straus. 2013. "Measuring quality of care: considering measurement frameworks and needs assessment to guide quality indicator development." *Journal of Clinical Epidemiology*, 66(12): 1320–1327.

Stokes, Tim. 2014. "QOF changes are ground-breaking and will need monitoring."

Strong, P, and J Robinson. 1990. *The NHS - Under New Management.* . 1st Paperback Printing edition ed., Milton Keynes England ; Philadelphia:Open University Press.

Stuart, Elizabeth A. 2010. "Matching methods for causal inference: A review and a look forward." *Statistical science : a review journal of the Institute of Mathematical Statistics*, 25(1): 1–21.

Sutcliffe, Daniel, Helen Lester, John Hutton, and Tim Stokes. 2012. "NICE and the Quality and Outcomes Framework (QOF) 2009-2011." *Quality in Primary Care*, 20(1): 47–55.

Taylor, Stephen James Lake. 1954. *Good General Practice. A report of a survey by Stephen Taylor.* Oxford University Press.

van de Vijsel, Aart R., Richard Heijink, and Maarten Schipper. 2015. "Has variation in length of stay in acute hospitals decreased? Analysing trends in the variation in LOS between and within Dutch hospitals." *BMC Health Services Research*, 15(1): 438.

Veninga, C. C. M, P Denig, R Zwaagstra, and F. M Haaijer-Ruskamp. 2000. "Improving drug treatment in general practice." *Journal of Clinical Epidemiology*, 53(7): 762–772.

Venkataramani, Atheendar S, Jacob Bor, and Anupam B Jena. 2016. "Regression discontinuity designs in healthcare research." *BMJ*, i1216.

Wachter, Robert M. 2013. "Personal accountability in healthcare: searching for the right balance." *BMJ Quality & Safety*, 22(2): 176–180.

Webb, R., and D. Hannay. 1996. "Career choices of trainees in general practice." *BMJ*, 312(7026): 314–314.

Wenger, Etienne. 1999. *Communities of Practice: Learning, Meaning, and Identity.* Cambridge University Press.

Wennberg, J., and null Gittelsohn. 1973. "Small area variations in health care delivery." *Science (New York, N.Y.)*, 182(4117): 1102–1108.

Wennberg, John, and Alan Gittelsohn. 1982. "Variations in Medical Care among Small Areas." *Scientific American*, 246(4): 120–135.

Wensing, M., T. van der Weijden, and R. Grol. 1998. "Implementing guidelines and innovations in general practice: which interventions are effective?" *British Journal of General Practice*, 48(427): 991–997.

Westert, Gert P., and Peter P. Groenewegen. 1999. "Medical practice variations: changing the theoretical approach." *Scandinavian Journal of Public Health*, 27(3): 173–180.

Westert, G.P. 1992. *Variation in use of hospital care.* Van Gorcum.

Westert, G. P., A. P. Nieboer, and P. P. Groenewegen. 1993. "Variation in duration of hospital stay between hospitals and between doctors within hospitals." *Social Science & Medicine (1982)*, 37(6): 833–839.

Wilkie, Veronica. 2014. "British general practice: another Collings moment?" *BMJ*, 349.

Williams, Simon J., Michael Calnan, Sarah L. Cant, and Joanne Coyle. 1993. "All change in the NHS? Implications of the NHS reforms for primary care prevention." *Sociology of Health and Illness*, 15(1): 43–67.

Wollersheim, H., R. Hermens, M. Hulscher, J. Braspenning, M. Ouwens, J. Schouten, H. Marres, R. Dijkstra, and R. Grol. 2007. "Clinical indicators: development and applications." *The Netherlands Journal of Medicine*, 65(1): 15–22.

Wordsworth, Sarah, Diane Skåtun, Anthony Scott, and Fiona French. 2004. "Preferences for general practice jobs: a survey of principals and sessional GPs." *The British Journal of General Practice*, 54(507): 740–746.

World Health Organization, OECD, and International Bank for Reconstruction and Development/The World Bank. 2018. *Delivering quality health services: a global imperative for universal health coverage.* World Health Organization.

Zhang, Zhongheng, Md Jamal Uddin, Jing Cheng, and Tao Huang. 2018. "Instrumental variable analysis in the presence of unmeasured confounding." *Annals of Translational Medicine*, 6(10).

Zohoori, Namvar, and David A. Savitz. 1997. "Econometric approaches to epidemiologic data: Relating endogeneity and unobserved heterogeneity to confounding." *Annals of Epidemiology*, 7(4): 251–257.