Continuity of Service Provision for Repeat Customers: Empirical Investigations of Continuity of Care in Primary Care Practices

Harshita Kajaria Montag
St. Edmund's College

This thesis is submitted for the degree of Doctor of Philosophy

September 2021

Judge Business School
University of Cambridge

Preface

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except in Chapters 2 and 4 as stated below. In Chapter 2, I collaborated with my supervisor, Professor Stefan Scholtes and co-author Professor Michael Freeman, and contributed 70% of the chapter. In Chapter 4, I collaborated with Professor Michael Freeman, and contributed 80% of the chapter. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in this Preface and specified in the text.

The research described herein was performed at the Judge Business School at the University of Cambridge between October 2017 and September 2021. It does not exceed the prescribed word limit for the relevant Degree Committee.

Harshita Kajaria Montag
Cambridge, UK
September 2021

Abstract

Name: Harshita Kajaria Montag
Title: Continuity of Service Provision for Repeat Customers: Empirical Investigations of Continuity of Care in Primary Care Practices

Many service processes are sequences of independent transactions between customers and servers. In such contexts it is well known that pooling server time and offering an arriving customer the next available server, independently of whether they had seen the customer before, is most efficient. However, when customer episodes depend on one-another, then this is not necessarily the case. Ensuring customers have a dedicated server may offer efficiency advantages. This thesis is an empirical investigation of the productivity advantages of dedicated servers in the context of primary care. In this context, server dedication equates with the notion of continuity of care, which is the extent to which a patient has her consultations with the same doctor. The thesis is composed of two in-depth empirical studies. The first study provides evidence of the productivity implications of relational continuity of care (RC), while the second study addresses the question how practice managers can increase continuity of care. Both studies use a detailed clinical dataset from the Clinical Practice Research Datalink that collects patient-consultation level electronic health records and is representative of the UK population. From its overall database of over 11 million patients spanning 674 practices, our data consists of the entire medical record of 5 million patients who have had contact with primary care between 2007 and 2018 across different providers of care (secondary care etc.). In the first study, I show that continuity of care has a significant productivity benefit. Specifically, due to the trust-based relationship, accumulated knowledge over time and stronger accountability for the patient, the average time to next visit for a patient is longer when patients see their regular provider than after they see another doctor. The data shows that the productivity benefit of care continuity is larger for older patients, patients with multiple chronic conditions, and patients with mental health conditions. I discuss operational and strategic implications of these findings for primary care practices in capitation environments and for third-party payers of their services. This first study suggests that prioritizing continuity of care is an effective strategy to reduce demand for consultations. The second study, entitled "The Operational Determinants of Relational Continuity of Care" builds on the first part and identifies the levers and strategies that primary care practice managers can implement to promote continuity within the constraint of a diminishing workforce. We find that a sustained increase in workload - caused by demand growth - and increasing fragmentation of the workforce - due to a shift to part-time and agency work - induces significant heterogeneity between practices in their ability to provide care continuity. Specifically, these two factors alone explain more than 50% of the decline in care continuity over the ten-year window of the study, with workforce fragmentation having the greater impact. We discuss the implications for workforce management in primary care practices that wish to promote continuity of care.

This thesis is dedicated to my parents,
for their endless love, support and encouragement

Acknowledgements

I am extremely grateful to my advisor Prof. Stefan Scholtes for his invaluable guidance and support. It is no exaggeration to say that this thesis wouldn't have been possible if not for his generosity, time, effort and patience. He has encouraged me to think independently, express my ideas confidently, and helped me develop my research style and acumen. He has always involved me in his engagements with industry experts and healthcare partners, helping me to build my own network. His candid advice has guided me throughout these past years, with Stefan finding the right words in every situation. I hope to continue to work with him and have plenty of opportunities to learn from him in the future.

I am also very fortunate to have been advised and mentored by Prof. Michael Freeman. He has been extremely generous with his time, always being available and patient to answer the smallest of queries, conducting brainstorming sessions, steering me through my first paper all the way through the publication process, and being open to any discussions about work or personal life. I am grateful for his continued support and guidance which I will surely need to pursue a successful career in academia.

I would also like to thank the faculty at the Judge Business School for creating an environment that fosters building the skill set necessary for working in academia. I am especially thankful to Prof. Paul Kattuman for allowing me to learn from and work with him on our joint project, as well as to Prof. Feryal Erhun and Prof. Houyuan Jiang (and the entire healthcare group) for their support and invaluable feedback on projects. In addition, I wish to thank our PhD directors Prof. Jennifer Howard-Grenville and Prof. Andrei Kirilenko, as well as our PhD program administrator Joanna Blakeman, without whom the PhD experience would not have been this seamless. I am also extremely grateful to Prof. Stelios Kavadias and Prof. Bradley Staats for their invaluable feedback on my thesis.

My thanks are extended to those with whom I have crossed paths with at Cambridge over the last four years. Two who deserve special mention are Katie, with whom I have enjoyed many conferences together and hope to continue to do so, and Lidia, whose friendship I will cherish forever. I am also grateful to my childhood friends Olivia and Abhilasha who have been my sounding board and uplifted me every time I needed them to.

Most importantly, I am blessed to have a loving and supporting family who have supported me at every step of the way. My husband, Alexander, is my pillar of strength and I am incredibly grateful to have travelled the doctorate journey alongside him. His unconditional love, support, patience, and wisdom has given me the strength to persevere in moments of doubt. Our daughter Aurelia is our biggest gift, who reminds us to be grateful for what we have and who keeps us grounded. My parents-in-law, Sabine and Andreas, along with their son Cornelius, welcomed me into their family in the warmest and kindest way and have given me a second home.

I thank my brother, Akhilesh, his wife Mudita, and their son, Shreyaan, in addition to my aunts, uncles and grandparents for their love and encouragement throughout this journey.

Finally, I am forever indebted to my parents, Anushree and Ravi, without whom none of this would have been possible and to whom I dedicate this thesis. They have always nurtured my desire to learn and supported all of my academic endeavours. Their unwavering belief in me has been my biggest source of motivation, and especially when the end often felt far from reach. Thank you for your endless and unconditional love, and for all the opportunities you have provided me to set me on the right path so that I could achieve anything I wanted to.

This work was supported by the Cambridge Judge Business School.

Exposure of this work

Under Review

Kajaria-Montag, H., M. Freeman, S. Scholtes. 2021. Continuity of care increases clinical productivity in primary care. *Major Revision at Management Science.*

Kajaria-Montag, H., M. Freeman. 2021. The Operational Determinants of Relational Continuity of Care: An Empirical Study of Primary Care in the UK. *Reject with an option to resubmit at Management Science.*

Conferences

Kajaria-Montag, H., M. Freeman, S. Scholtes, Continuity of care increases clinical productivity in primary care. *INFORMS Annual Meeting*, Seattle, Washington, USA. October 2019.

Kajaria-Montag, H., M. Freeman, The Operational Determinants of Relational Continuity of Care: An Empirical Study of Primary Care in the UK. *INFORMS Annual Meeting*, Virtual. October 2020.

Kajaria-Montag, H., M. Freeman, The Operational Determinants of Relational Continuity of Care: An Empirical Study of Primary Care in the UK. *POMS 2021 Annual Conference*, Virtual. April 2021.

Kajaria-Montag, H., M. Freeman, The Operational Determinants of Relational Continuity of Care: An Empirical Study of Primary Care in the UK. *INFORMS MSOM Healthcare Special Interest Group*, Virtual. June 2021.

Kajaria-Montag, H., M. Freeman, S. Scholtes, Continuity of care increases clinical productivity in primary care. *INFORMS MSOM Conference*, Virtual. June 2021.

Kajaria-Montag, H., M. Freeman, S. Scholtes, Continuity of care increases clinical productivity in primary care. *INFORMS Healthcare*, Virtual. July 2021.

Kajaria-Montag, H., M. Freeman, The Operational Determinants of Relational Continuity of Care: An Empirical Study of Primary Care in the UK. *INFORMS Annual Meeting*, Hybrid. October 2021.

Kajaria-Montag, H., M. Freeman, **S. Scholtes**, Continuity of care increases clinical productivity in primary care. *INFORMS Annual Meeting*, Hybrid. October 2021.

Kajaria-Montag, H., M. Freeman, S. Scholtes, Continuity of care increases clinical productivity in primary care. *POMS Annual Meeting*, Virtual. April 2022.

*Presenting author highlighted in **bold** font.*

Contents

Chapter 1

Introduction

Repeat interactions between a server and the customer is an important metric to assess the health or success of a service. Strategic service providers spend significant amount of time and money to maintain good, long-term relationships with their customers and are eager to keep their customers satisfied, often with different goals depending on the service context. Most customers also want to interact and use the services of someone they know and can trust, and wish to avoid having to spend time searching for alternative options.

The notion of the customer-server relational continuity is akin to the role of the dedicated client relationship manager, whose goal is to build a relationship with the client which is based on trust and value, and the longevity of the relationship depends on the quality of the interactions and the accumulated knowledge over time. These client relationship managers are prevalent in various knowledge intensive industries such as banking and healthcare. At banks, these services are provided for wealthy individuals whose private wealth managers help them plan their monetary future, diversify their investment portfolios, provide them with the knowledge, information, experience, and skills to sustain and expand their capital. In healthcare, having a relationship manager is comparable to having a dedicated physician who is holistically accountable for the care of the patient throughout, and is the physician the patient sees every time they visit.

In operations management, repeated interactions have been studied in the context of interactions with a 'task' or 'platform' which does not account for human-to-human interactions. With the rise in work that involves interaction between people, relationships warrants further study.

Studies in revenue management have looked at the interactions between the customer and the platform in the backdrop of a dynamic repeated setting rather than a one-shot transaction, and how the goodwill of the customer towards the platform can change over time. This has consequences on the optimal service rates and the optimal revenue maximization policy (Calmon et al. 2021). Popescu and Wu (2007) look at a dynamic pricing problem where repeated interactions leads customers to anchor on a certain price reference, and any changes in prices are seen as surcharges or discount, which, according to the frequency of the customer-service interaction and the loss aversion degree of the customer, has long-term implications for the firms revenue.

Loyalty programs, which are traditionally marketing tools used mostly in the retail and airline industries, are also beginning to be integrated into the revenue management framework. The concept underpinning loyalty programs is in fact the importance of repeated interactions with customers and is an important lever for generating consumer demand and the firms optimal

pricing policies for different customer segments, making it a key operational decision for the firm (Chun and Ovchinnikov 2019, Chun et al. 2020).

In a recent paper by Clark et al. (2013) show that in outsourcing settings, focal customer experience (interaction between a server and a focal customer) is an important determinant of learning in service contexts. Specifically customers and servers interact to co-produce the service output, and the improved communication gained from repeated interactions leads to better performance (as defined by length of time to complete a task). Though, this study is unable to examine the impact of focal customer experience on any quality measure. My setting in this study enables me to study the impact of such focal customer experience on quality.

In many service contexts, it is common to think of the throughput or server productivity as number of visits per day for a given server. Such a metric would guarantee, based on the traditional service literature, lower costs, and more effective satisfaction of the demand. But, in service settings where visits are interdependent, I suggest that a novel way of thinking about productivity, which is defined as the frequency of returning "customers", as captured by the time between visits from a particular customer. Though this sounds like an alternative way to capture the traditional productivity measure, I argue that the eventual workload faced by the firm is affected by the characteristics of the visit itself. Depending on these characteristics, the returning frequency might be lower (i.e., longer time between visits) and the system can cope with the demand while delivering good service. Moreover, this measure has the advantage of being a quality metric – if servers sort out customers 'well', they do not need to return as frequently. The specific visit characteristic I look at in this thesis is 'who' performs the service, or in other words, the 'level of dedication' to the customer. A dedicated server might reduce the effectiveness of pooling that is often considered more efficient by service system experts, but the benefits of higher dedication can supersede the inefficiencies from lack of pooling. Given the advantages of the dedicated relationships, and the parallel rise of the gig economy, in order to deliver dedicated service, organizations must maintain a core established workforce rather than rely on a temporary workforce. In the second part of my thesis, I explore the erosion of a dedicated service due to workforce fragmentation.

Below, I outline the literature streams that are relevant to the thesis and how the findings of the thesis contribute to each of these literature streams.

Pooled vs Dedicated queues

In the operations management literature, providing customer-server continuity as akin to the notion of dedicated queues as opposed to pooled queuing systems. Pooled queues are generally advocated due to the efficiency advantages that can be gained when the service times are independently and identically distributed. Though, various studies have also discussed situations when dedicated queues are preferred, some of these include – when customers have heterogeneous needs and consequently highly variable service times (Smith and Whitt (1981), Benjaafar

(1995)), when customers are delay-sensitive (Sunar et al. (2021)), when servers are non-identical (Smith and Whitt (1981), Rubinovitch (1985)), and when servers in a pooled setting might exhibit 'social loafing' (Wang and Zhou (2018)). We add to this growing body of literature by showing that in knowledge-intensive settings such as the one I consider, dedicated queues provide information advantages to a server through which the server can provide a more appropriate and customized service to the customer, and consequently reduce future demand.

Gatekeeping literature

The studies also provide insights for gatekeeping settings especially in the context of how the service decisions are affected by congestion. In gatekeeping settings, customers interact with a server who is a gatekeeper; the gatekeeper either provides the service himself or refers the customer to a specialist.The decision whether to self-serve or refer is based on the trade-off between protecting the valuable time of the specialist or potentially failing to resolve the customers problem (the probability of which increases in the complexity of the customer) (Shumsky and Pinker 2003). Moreover, studies show that an increase in workload (measured as number of customers per server) leads to higher variability in customer service experience, specifically, a higher gatekeeper referral rate for complex customers and a lower resource intensive service for non-complex cases (Freeman et al. 2020). The results of this thesis have implications for this stream of literature related to workload management in gatekeeper settings, specifically, if there is gatekeeper-customer continuity, especially for the more complex patients, the gatekeeper will be able to provide better service to the customer and smooth the future gatekeeper workload, both of which will enable the protection of the more expensive downstream services.

Organisational learning

The thesis borrows from the organization learning literature to show the extent to which organizations maintain a core established workforce varies and consequently, the collective rates of dedicated service or continuity that they are able to provide. While freelancers or temporary workers provides flexibility in scheduling practices, it makes it more difficult for the customers to access their regular dedicated servers. Additionally, the performance of the freelance workers, in terms of quality, might be lower than the core established workforce. Studies focusing on organizational learning effects can provide insight into why this might be the case.

The drivers of different learning rates across organizations are not only affected by the rate of cumulative individual experience, but also the ability of individuals to leverage the knowledge of other team members, as well as the capacity of the practice to be coordinated (Reagans et al. 2005). As a consequence, some practices might have dramatic learning rates. These characteristics of the practice might lead the workers to develop organization-specific skills and gain familiarity over time (Huckman and Pisano 2006). This phenomena is particularly pronounced

when workers, such as surgeons perform operations at multiple hospitals or freelancers such as locums spend time at different primary care practices. Their performance seems to improve significantly with an increase in volume of the specific task at that given organization (Huckman and Pisano 2006). Similarly, evidence shows that for a particular surgeon's experience at a focal task is also firm-specific, and yields better surgeon performance than non-focal experience, or a portfolio of experience (Kc and Staats 2012), which is particularly concerning as more and more primary care doctors have been taking on non-clinical tasks (Baird and Holmes 2019). Moreover, if the focal task is specific to the customer, it can also be benefical for the individual learning of the worker (Clark et al. 2013). For a specific task, firms can also improve efficiency and coordination by formally prioritizing their goals, improving communication and consistency, enhancing team memerbship and having a consistent group of people working on the same task (Pisano et al. 2001).

Hence, if we apply organizational learning to the goal of providing continuity, we find that organizations that are able to foster a better learning environment, retain a core workforce and formally prioritize continuity in a coordinated manner, are able to offer a higher level of continuity of care to their customers.

Workforce literature.

Various studies in the management literature have focused their attention on the workforce of the organization or the nature of servers in the context of team composition and team familiarity. Specifically these studies look at whether to have diverse or specialized teams (Glover and Kim 2021, Hoogendoorn et al. 2017), and the degree of familiarity and exposure amongst team members and its effect on performance (Huckman and Staats 2011). Though these studies provide valuable lessons in team dynamics for our context of doctors working at the same primary care practice and being part of a larger team, the doctors maintain individual relationships with patients. Moreover, the doctors are accountable for the patients on their panel and therefore continuity of care is not necessarily team-based.

Other workforce management literature includes scheduling decisions (Bard and Wan 2008), workforce agility (Iravani and Krishnamurthy 2007) and benefits of cross-trained workers (Pinker and Shumsky 2000), but these are studied in settings that are not knowledge intensive nor have repeated interactions with the same customers, which forms the premise of our study.

Some studies have looked at the use of flexible labor resources, or the optimal use of both part-time and full-time workforce (Dong and Ibrahim 2020). Similar to our study is the study by Kesavan et al. (2014) who find that use of part time workers to cope with demand has an inverted U-shaped relationship with financial performance, where too much use of part time workers has negative consequences for performance.

Unlike this existing work on workforce management, our study looks at workforce composition in

a knowledge intensive setting with repeated interactions, combining different aspects of workforce composition, such as flexible labor composition (part-time vs full time), hierarchical composition (partners vs. non-partners), resource intensity based composition (doctors vs. nurses) and contract-related composition (permanent vs temporary).

In my thesis, I combine the notion of repeated interactions, as well as the trust and accumulated knowledge gain from a long-term interpersonal relationship between a server and a client, and look at how capitalizing on this interdependence of customer episodes can lead to better outcomes. The focus is on the knowledge-intensive healthcare industry, particularly in the primary care setting and the interaction between the primary care physician and the patient. It is integral to understand what characterizes the patient's experience of seeing the same doctor that makes it desirable to offer continuity. The relationship between the patient and doctor is one that is built on understanding and trust. Through repeated interactions, the doctor builds knowledge of the patient and patients' condition, and forms a trust-based relationship that may help her to communicate more effectively with the patient (Reagans et al. 2005). As a consequence, the doctor may have a more complete understanding of the available treatment options, make a more appropriate choice of treatment plan, and guarantee better compliance with the plan. The continuity doctor feels more accountable is more likely to spend extra time or effort with the patient to "get it right the first time" instead of doing the minimum necessary to alleviate symptoms.

Though, as the workforce is shrinking and unable to cope with rising demand, continuity is being less and less prioritized, and consequently the quality of interactions is suffering. This has also led to the advent of transactional platforms such as AmazonCare and Babylon, which are shifting the nature of primary care models to provide one-time services to patients with doctors they do not have a relationship with. Considering the many advantages of continuity for patients, these shifts have led us to think about novel ways of offering care in the constrained setting, whether continuity is still beneficial in the primary care setting and who it is beneficial for. This might require restructuring and rethinking of the service offering to provide continuity to a segment of the population through traditional primary care practices, while offering transactional services in a platform-based environment to the rest.

The thesis comprises of two empirical studies that helps us to better understand the implications of providing service continuity in the primary care setting itself – particularly, in contrast to belief that practices must trade-off productivity and care continuity as competing goals, we find that continuity in fact enhances clinical productivity – and secondly, knowing that there are untapped productivity gains to be had from continuity, how can the primary care practice manager improve their provision of service continuity.

In the rest of the introduction we provide a more specific discussion of the healthcare context that motivates these works, together with a preview of the headline findings.

1.0.1 Healthcare setting

Primary care is facing a workforce crisis. Forecasts suggest that there will be a shortage of up to 55,000 primary care physicians (PCPs) in the United States (US) by 2033 (Heiser 2019), while the United Kingdom (UK) was estimated to have 2,500 fewer full-time equivalent (FTE) general practitioners (GPs) than needed in 2019 with this shortage projected to almost triple by 2024 (Beech et al. 2020). These trends pose a critical challenge for primary care managers: how can a stagnant or shrinking clinical workforce manage the increasing demand for primary care services? Most primary care practices have sought to meet this challenge and satisfy the rising demand for consultations by improving daily throughput and offering more consultations per clinician day. While these actions appear to have been partially effective in combating the growing demand for consultations, they also seem to have taken a toll on the workforce. Stress, absenteeism, and burnout are endemic in the industry, driving many clinicians to leave the profession or transition to part-time or more flexible working practices (The King's Fund 2019). Moreover, recent trends indicate that waiting times for primary care are on the rise and patient satisfaction is declining, suggesting that such strategies may no longer be sufficient on their own (Thorlby et al. 2019).

In light of these challenges, primary care practices might consider adopting another approach to mitigate the growing workload and workforce pressures: reducing demand by keeping their patients as healthy as possible and pro-actively reducing their need for consultations. But for a manager of a primary care practice, executing on this approach may seem less straightforward. With the system already strained, it is not immediately clear how to improve patient outcomes without providing patients with more care, at least in the short term.

However, the first study in the thesis suggests that a potential solution may be to not provide patients with more care but instead focus on who is providing that care. In particular, we show that when a patient is seen by the doctor with whom they are most familiar, they have shorter consultation times and return later for a subsequent visit. The effect is more pronounced for more comorbid patients, older patients and patients suffering from mental health conditions. This finding suggests that practice managers who prioritize relational continuity of care (RC) may be more effective at reducing demand for consultations and improving the productivity of their clinical workforce.

Importantly, RC not only has the potential to improve clinical productivity but also has been shown to drive better patient health outcomes (e.g., lower mortality rates) and improve system performance (e.g., reduce ED visits). These advantages are perhaps not surprising since, in contrast to the episode-focused secondary care model in which "diseases stay and patients come and go," in the primary care setting "patients stay and diseases come and go" (Heath 1995). Repeated interactions can thus increase GP's sense of ownership and personal responsibility for their patient's health and well-being, improving clinical decision making and job satisfaction (Grembowski et al. 2005). Meanwhile, with health systems and governments increasingly

trialing and adopting integrated care models to reduce costs and improve efficiency, primary care practices that are better at providing RC may find in the future that they are financially rewarded for better patient outcomes and system performance. Given the various potential benefits of RC for patients, providers, and health systems, the central question of the second study is, therefore: what are the levers available to a primary care practice manager looking to deliver RC at their practice?

It turns out that the answer to this question is not so straightforward. Surprisingly, there is an absence of practitioner-oriented advice and a lack of evidence-based academic research to inform strategies to promote RC. A 2010 evidence review by The King's Fund, a UK-based health policy think-tank, for example, concluded: "We were struck by the absence of agreed policies or any general body of expertise on how to encourage continuity. Specific guidance is also lacking from the Royal College of General Practitioners. Meanwhile, many developments in practice and national policy have had the unintended consequence of making relationship continuity more difficult to achieve" (Boyle et al. 2010). Although practitioners, professional bodies, and policymakers have shown a renewed interest in RC of late, there is still an absence of evidence as to which factors affect a practice's ability to provide RC (Jeffers and Baker 2016).

Thus, the second part of the thesis sets out to fill this gap in the literature by identifying strategies that primary care practice managers can implement to promote continuity within the constraint of a diminishing workforce. This paper provides the first empirical study on the antecedents of RC in the primary care context. Using the insights from this analysis, practice managers can identify the root cause of low rates of RC at their own practice and the key operational levers that they can use to promote continuity. We find that a sustained increase in workload - caused by demand growth - and increasing fragmentation of the workforce - due to a shift to part-time and agency work - induces significant heterogeneity between practices in their ability to provide continuity. In fact, these factors alone can also explain more than 50% of the decline in RC over the past decade, with workforce fragmentation having a relatively greater impact than demand growth.

References

Baird B, Holmes J (2019) Why can't I get a doctor's appointment? Technical report, The King's Fund, URL https://www.kingsfund.org.uk/publications/solving-issue-gp-access.

Bard JF, Wan L (2008) Workforce design with movement restrictions between workstation groups. *Manufacturing & Service Operations Management* 10(1):24–42.

Beech J, Bottery S, Charlesworth A, Evans H, Gershlick B, Hemmings N, Imison C, Kahtan P, McKenna H, Murray R, Palmer B (2020) Closing the gap: Key areas for action on the health and care workforce — The Nuffield Trust. Technical report, Nuffield Trust.

Benjaafar S (1995) Performance bounds for the effectiveness of pooling in multi-processing systems. *European Journal of Operational Research* 87(2):375–388.

Boyle S, Appleby J, Harrison A (2010) A rapid view of access to care. Technical report, The Kings Fund.

Calmon AP, Ciocan FD, Romero G (2021) Revenue management with repeated customer interactions. *Management Science* 67(5):2944–2963.

Chun SY, Iancu DA, Trichakis N (2020) Loyalty program liabilities and point values. *Manufacturing & Service Operations Management* 22(2):257–272.

Chun SY, Ovchinnikov A (2019) Strategic consumers, revenue management, and the design of loyalty programs. *Management Science* 65(9):3969–3987.

Clark JR, Huckman RS, Staats BR (2013) Learning from customers: Individual and organizational effects in outsourced radiological services. *Organization Science* 24(5):1539–1557.

Dong J, Ibrahim R (2020) Managing supply in the on-demand economy: Flexible workers, full-time employees, or both? *Operations Research* 68(4):1238–1264.

Freeman M, Robinson S, Scholtes S (2020) Gatekeeping, fast and slow: An empirical study of referral errors in the emergency department. *Management Science* .

Glover J, Kim E (2021) Optimal team composition: Diversity to foster implicit team incentives. *Management Science* .

Grembowski D, Paschane D, Diehr P, Katon W, Martin D, Patrick DL (2005) Managed care, physician job satisfaction, and the quality of primary care. *Journal of General Internal Medicine* 20(3):271–277.

Heath I (1995) Fortnightly Review: Commentary: The perils of checklist medicine. *Bmj* 311(7001):373, ISSN 14685833, URL http://dx.doi.org/10.1136/bmj.311.7001.373.

Heiser S (2019) New Findings Confirm Predictions on Physician Shortage. *AAMC News* 1–3.

Hoogendoorn S, Parker SC, Van Praag M (2017) Smart or diverse start-up teams? evidence from a field experiment. *Organization Science* 28(6):1010–1028.

Huckman RS, Pisano GP (2006) The firm specificity of individual performance: Evidence from cardiac surgery. *Management Science* 52(4):473–488.

Huckman RS, Staats BR (2011) Fluid tasks and fluid teams: The impact of diversity in experience and team familiarity on team performance. *Manufacturing & Service Operations Management* 13(3):310–328.

Iravani SM, Krishnamurthy V (2007) Workforce agility in repair and maintenance environments. *Manufacturing & Service Operations Management* 9(2):168–184.

Jeffers H, Baker M (2016) Continuity of care: still important in modern-day general practice.

Kc DS, Staats BR (2012) Accumulating a portfolio of experience: The effect of focal and related experience on surgeon performance. *Manufacturing & Service Operations Management* 14(4):618–633.

Kesavan S, Staats BR, Gilland W (2014) Volume flexibility in services: The costs and benefits of flexible labor resources. *Management Science* 60(8):1884–1906.

Pinker EJ, Shumsky RA (2000) The efficiency-quality trade-off of cross-trained workers. *Manufacturing & Service Operations Management* 2(1):32–48.

Pisano GP, Bohmer RM, Edmondson AC (2001) Organizational differences in rates of learning: Evidence from the adoption of minimally invasive cardiac surgery. *Management Science* 47(6):752–768.

Popescu I, Wu Y (2007) Dynamic pricing strategies with reference effects. *Operations research* 55(3):413–429.

Reagans R, Argote L, Brooks D (2005) Individual experience and experience working together: Predicting learning rates from knowing who knows what and knowing how to work together. *Management science* 51(6):869–881.

Rubinovitch M (1985) The slow server problem. *Journal of Applied Probability* 205–213.

Shumsky R, Pinker E (2003) Gatekeepers and referrals in services. *Management Science* 49(7):839–856.

Smith DR, Whitt W (1981) Resource sharing for efficiency in traffic systems. *Bell System Technical Journal* 60(1):39–55.

Sunar N, Tu Y, Ziya S (2021) Pooled vs. dedicated queues when customers are delay-sensitive. *Management Science* .

The King's Fund (2019) The future of general practice — The King's Fund. Technical report, The King's Fund.

Thorlby R, Gardner T, Turton C (2019) Nhs performance and waiting times. *The Health Foundation* .

Wang J, Zhou YP (2018) Impact of queue configuration on service time: Evidence from a supermarket. *Management Science* 64(7):3055–3075.

Chapter 2

Continuity of care increases clinical productivity in primary care

Continuity of care, defined as an ongoing therapeutic relationship between a patient and a physician, is a defining characteristic of primary care services. However, arranging a consultation with one's regular doctor is increasingly difficult as practices seek to improve daily throughput to match rising consultation demand with an increasingly scarce supply of clinical labor. The emergence of online providers accelerates this trend. We study the productivity implications of this reduction in care continuity by analyzing consultation-level data of over 10 million office consultations in 381 English primary care practices over a period of 11 years. We find that continuity of care has a significant productivity benefit. On average, the time to a patient's next visit is 13.2% (95% CI: $[12.2\%, 14.1\%]$) longer when the patient sees the doctor they have seen most frequently over the past two years, while there is no operationally meaningful difference in consultation length. The results are consistent across several model specifications that account for confounding and selection bias. The data shows that the productivity benefit of care continuity is larger for older patients, patients with multiple chronic conditions, and patients with mental health conditions. We estimate that the total consultation demand in our sample would have fallen by 5.2% had all practices offered continuity of care at the level of the top decile of practices while prioritizing patient consultations expected to yield the largest productivity benefits. We discuss operational and strategic implications of these findings for primary care practices and for third-party payers of their services.

2.1 Introduction

Primary care practices around the world are experiencing rising demand at a time when their most critical resource, primary care physician time, is becoming more scarce and more expensive. In the UK, the number of primary care physicians per 100,000 population decreased from 67 in 2009 to 60 in 2018, despite increasing demand from an aging population (Palmer 2019). The UK's Nuffield Trust estimates a shortfall of 7,000 general practitioners in the UK by 2023-24 (Beech et al. 2020), while the Association of American Medical Colleges (AAMC) estimates a shortfall of between 21,400 and 55,200 primary care physicians in the US by 2033 (Dall et al. 2020). These projections were made before the COVID-19 pandemic, and the additional pressure

caused by COVID-19 will likely exacerbate these shortfalls. In response, primary care practices need to increase their clinical productivity.

Within a primary care practice, clinical productivity has two main dimensions: the number of consultations that a clinician performs in a day and the extent to which a clinician is able to extend the time to a patient's next consultation. Much emphasis is currently put on the former – increasing daily throughput – not least because it directly improves on-the-day access for patients. This focus on daily throughput has an unintended but important consequence: It becomes more difficult for patients to arrange a consultation with their regular doctor (Kajaria-Montag and Freeman 2020). In 2009, 77% of respondents to the UK's annual GP Practice Survey reported being able to see their preferred doctor at least most of the time. Ten years later, in the last survey prior to COVID-19, this proportion had dropped to 45% (Institute for Government 2019). Continuity of care is at danger of becoming an after-thought, a nice-to-have, rather than an essential feature of a low-cost primary care model.

Traditionally, general practitioners and their patients have valued the sustained, trust-based therapeutic relationships they form over time (Liu et al. 2018). Through repeated interactions, family doctors become familiar with the patient's holistic medical needs as well as with their preferences, behaviors, and socioeconomic circumstances, and will customize their advice accordingly. Continuity of service in primary care is also known to be beneficial for both patients and the wider health system. A recent evidence review by the UK's Nuffield Trust concluded that "relational continuity of care in general practice is associated with [...] better clinical outcomes for an array of conditions; reduced mortality; better uptake of preventative services; better adherence to medication; reduced avoidable hospital admissions; and better overall experience of care among patients who prefer continuity and are able to obtain it" (Palmer et al. 2018).

Given the strong evidence of its patient health and system benefits, it is surprising that relatively little is known about the effect of care continuity on the productivity of primary care practices themselves. It is particularly important to fill this gap in the literature given the current trends in primary care provision. Many primary care practices are responding to the workforce crisis by doubling down on daily throughput improvements, e.g., through standardizing and shortening the length of patient consultations or grouping physicians into rapid access teams that serve on-the-day patient demand in a round-robin fashion akin to an emergency department. This makes care continuity harder to achieve (Sampson et al. 2008), suggesting that practices must trade off clinical productivity and care continuity as competing goals.

Yet our analysis shows that this trade-off is illusory. In particular, our data indicates that care continuity *increases* a physician's productivity by extending the interval between patient visits. Throughput-enhancing interventions will therefore have unintended productivity-reducing consequences if they decrease continuity of care. From a more positive perspective, this also means that there are untapped productivity gains to be had from a targeted increase in continuity of

care.

To empirically demonstrate the productivity implications of care continuity, we use a sample containing over 10 million face-to-face consultations between over 14,000 primary care physicians and 1.8 million patients in 381 English primary care practices over a period of 11 years. For each consultation, we identify the patient's regular doctor as the doctor who had the most frequent interactions with the patient over the past two years. We then analyze whether a patient's revisit interval (i.e., the time between the focal consultation and the patient's next visit) differed when the consultation was with this regular doctor or with another doctor in the practice. Using a range of empirical methods to control for potential selection and omitted variable bias, we find robust evidence that the revisit interval is extended by 13.2% (95% CI: $[12.2\%, 14.1\%]$) if the patient sees her regular doctor. At the same time, we find no evidence that the patient's consultations with the regular doctor are longer than her consultations with other doctors in the practice. In fact, the data suggests that the consultation time with the regular doctor is shorter on average. Therefore, care continuity has a net positive overall effect on a primary care physician's productivity because it substantially extends the average time to the patient's next visit.

Having established the main effect, we then analyze the heterogeneity of the effect and show that the productivity benefit of care continuity is more pronounced for patients with more complex needs, specifically older patients, patients with chronic diseases and patients with mental health conditions. Building on these findings, we demonstrate how our estimation methods can be used as a scoring tool to enable practice managers to target relational services at those patients for whom care continuity has the most productivity-enhancing effect. We apply this scoring method retrospectively to the data to estimate the potential for demand reduction. The model suggests that if all practices in the data had offered well-targeted continuity of care for 75% of their consultations, a level similar to the 90th percentile across all practices in the data, then the overall demand for consultations would have been 5.2% lower than the realized demand.

2.2 Literature Review

This is the first study to our knowledge that considers the relationship between service continuity and productivity. Within the wider service management literature, the impact of various interventions or characteristics on worker productivity has been well documented across different settings. Some examples of studies relating to worker productivity include: impact of IT monitoring on employee productivity in restaurants (Pierce et al. 2015), the effect of worker interruption in terms of machine breakdowns on worker productivity (Cai et al. 2018), impact of public relative performance on workers in the hospital (Song et al. 2018), effect of manager gender on female worker productivity in a garment factory (Ranganathan and Shivaram 2021), impact of workload on productivity in a restaurant chain (Tan and Netessine 2014), the impact

of corporate wellness programs on productivity in laundry plants (Gubler et al. 2018), and the impact of specialization versus variety on productivity in the banking industry (Staats and Gino 2012).

It is common to think of the throughput or server productivity as number of visits per day or hour for a given server. Such a metric would guarantee, based on the traditional service literature, lower costs, and more effective satisfaction of the demand. But, in service settings where visits are interdependent, I suggest a novel way of thinking about productivity: the frequency of returning "customers", as captured by the time between visits from a particular customer. Moreover, this measure has the advantage of being a quality metric – if servers sort out customers 'well', they do not need to return as frequently. If the service is delivered well, returning frequency might be lower (i.e., longer time between visits) and the system can cope with the demand while delivering good service

I argue that how 'well' the customers are sorted out in during the visit is affected by the relationship between the customer and the server. Specifically, I look at is 'who' performs the service, or in other words, the 'level of dedication' to the customer. Below we highlight some examples of studies that have looked at the advantages of relationships. Within operations, the importance of customer-server repeated interactions and relationship is well highlighted in the context of contracting between a firm and a supplier, which is especially prevalent in the automobile, semiconductor and electronics industry. When the parties are not yet at the stage of writing court-enforceable contracts, repeated interactions (or anticipated repeated interactions) encourage the adoption of relational contracts (informal agreements) that are built on trust and cooperation (Taylor and Plambeck 2007). The degree of how much capacity the supplier is willing to set aside for the particular buyer is dependent on the strength of the relationship between them.

Studies in revenue management have looked at the interactions between the customer and the platform in the backdrop of a dynamic repeated setting rather than a one-shot transaction, and how the goodwill of the customer towards the platform can change over time. This has consequences on the optimal service rates and the optimal revenue maximization policy (Calmon et al. 2021). Popescu and Wu (2007) look at a dynamic pricing problem where repeated interactions leads customers to anchor on a certain price reference, and any changes in prices are seen as surcharges or discount, which, according to the frequency of the customer-service interaction and the loss aversion degree of the customer, has long-term implications for the firms revenue.

Loyalty programs, which are traditionally marketing tools used mostly in the retail and airline industries, are also beginning to be integrated into the revenue management framework. The concept underpinning loyalty programs is in fact the importance of repeated interactions with customers and is an important lever for generating consumer demand and the firms optimal pricing policies for different customer segments, making it a key operational decision for the firm (Chun and Ovchinnikov 2019, Chun et al. 2020).

Evidence from financial services suggests that since it is costly to search for service providers,

repeated interactions between investment banks (the provider) and investors (the customer) can lead to favorable pricing of convertible bonds (Henderson and Tookes 2012). Repeated interactions are also beneficial for bank loan provision and more favorable IPO allocations (Diamond 1991, Reuter 2006). For small firms, a study in the management literature finds that relational trust between banks and small firms mediates the relationship between bank strategies (such as manager continuity with the small firm) and customer loyalty (Saparito et al. 2004).

These studies consider relationships between organizations or between customers and a platform, but do not consider relationships between people. This study specifically considers how the patient-doctor relationship has advantages not only for health outcomes but in fact has implications for the future workload and productivity for the organization as well.

This paper contributes to the healthcare operations and medical literature on continuity of care. These literature streams have largely focused on the consequences of care continuity on patient outcomes and secondary care utilization, such as emergency visits or hospital admissions, typically for patients with specific conditions. In contrast, we are concerned with the effect of care continuity on the clinical productivity of primary care practices themselves, averaged across all patient consultations as well as for patients with specific conditions. Our insights are also relevant to the operations management literature on queuing systems, specifically the benefits of dedicated queue configurations. In this section, we highlight how our study is positioned within these literature streams.

2.2.1 Continuity of care in healthcare operations

Ahuja et al. (2020b) investigate the effect of providing continuity to patients with diabetes and find that continuity improves three important system utilization metrics: inpatient admissions, hospital length of stay, and readmission rate. In a follow-up study, Ahuja et al. (2020a) partially explain this relationship between continuity and the system utilization metrics by showing that continuity can lead to higher rates of medication adherence and consequently to lower glycemic variability. Senot (2019) also studies the effect of continuity on secondary care usage. Specifically, the study follows the journey of heart failure patients over a one-year period and finds that the continuity of the individual referring provider (along with continuity of the physical location and the accountable care organization) contributes to a reduced risk of hospital readmission for heart failure patients. Queenan et al. (2019) find that providing technology-enabled continuity coupled with increasing patient engagement in their own health reduces hospital readmissions.

The present study complements this research on system utilization effects by focusing on productivity effects within the primary care setting itself. This is important because the direction of the internal productivity effect tells us whether or not practices need to be externally incentivized to provide the continuity of care that will create the documented system utilization benefits.

The healthcare operations literature has hitherto not engaged much with primary care productivity, though Bavafa et al. (2018) is one exception. Their paper focuses on the impact of complementing office visits by e-visits on demand in primary care, and the authors show that the introduction of this new communications channel increases demand for office visits. We also address how demand for primary care services changes as a function of the service provided, with our focus being the demand-inducing effect of reduced care continuity. While Li et al. (2021) also focus on telemedicine adoption, they do so in an outpatient context. They show that the adoption of telemedicine reduces productivity by shortening the interval between patient visits in the short term, but the interval between visits increases in the long run. We also use the revisit interval as a measure of productivity in this paper, but our focus is on the productivity implication of care continuity in primary care.

Another study that is closely related to this paper is Song et al. (2015). The authors empirically investigate the effect of allocating responsibility for patient-flow decisions and medical decisions for a patient to the same clinical decision-maker, rather than separating responsibility for the two functions. They show that introducing a single decision-maker leads to a significant reduction in length of stay in the ED. The authors attribute this to increased ownership over patient flow and a better alignment of incentives (i.e., physicians in the dedicated system can only leave their shift when all their allocated patients are discharged) – ideas explored further in Armony et al. (2021), a modeling follow-up paper. Our paper borrows from Song et al. (2015) the theoretical lenses of "ownership" and "incentive alignment" when developing the hypotheses. However, our study differs in important aspects. First, due to repeated patient-provider interactions in primary care, a patient's preferred provider has ownership of the patient's long-term health (rather than a single health episode, as is commonly the case in the ED context). Second, because of these repeated interactions, the preferred clinician is incentivized to "get it right first time" to avoid a revisit – meaning that they may instead opt to spend longer with each patient. Finally, pure throughput metrics, such as length of stay or number of appointments per day, are inappropriate in a context with repeated interactions. Therefore, we focus on prolonging revisit intervals, which is likely to signal both a productivity and a quality benefit.

2.2.1.1 Dedicates queues and revisit interval

In the context of our study, providing continuity to patients, or allocating a patient to his preferred provider is similar to setting up a dedicated queue configuration.

In queueing literature, although it is widely shown that pooled service configurations reduce long run average times, there have been significant number of studies that demonstrate when dedicated queues might be more efficient. From the *customer's* perspectives, dedicated queue configurations provide significant gains over pooled queues specifically when customers are heterogeneous (Benjaafar 1995), when customers jockey between separate queues (Rothkopf and Rech 1987), or when they are delay-sensitive (Sunar et al. 2021). From the perspective of the

server, studies show that when the servers are non-identical (Smith and Whitt 1981, Rubinovitch 1985), dedicated queue configurations are preferred. Other than customer or server heterogeneity, other mechanisms might be at play which make dedicated queues more desirable in certain settings. For example, in a call centre setting, where customers and servers are assumed to be homogenous, (Jouini et al. 2008) show that when customers are grouped and served by a dedicated team of agents, the smaller size of the individual groups can enable easier management of the customers, hence improving speed and quality.

Empirical studies investigating the efficiency of different queue configurations, have also demonstrated the benefits of dedicated queues. (Wang and Zhou 2018) show that in the supermarket setting, servers in a pooled might exhibit "social loafing" when they slow down to reduce the work and effort they themselves have to exert, hence making the service times for dedicated queues more efficient. (Song et al. 2015) show that in the ED setting, dedicated queues decrease LOS and average wait time of the patients via the mechanism of physician ownership which is enabled by the physicians have more control by being able to manage the flow of patients in and out of ED beds.

In contrast to these server-driver mechanisms, we look at a setting in which both the servers and the customers are heterogeneous and the benefits of dedicated queues are derived from the relationship between the customer and server. More specifically, our setting combines both patient and physician related mechanisms by looking at the relationship between them. Through repeated interactions, the doctor builds knowledge of the patient and patients' condition, and forms a trust-based relationship that may help her to communicate more effectively with the patient (Reagans et al. 2005). Physicians develop a sense of ownership and accountability with less incentive to "pass on the buck" to someone else, patients confide in their physicians by discussing their personal and social needs which improves adherence and compliance, and overall the physicians have a more holistic understanding of the patients needs.

We hypothesize that these aspects of the patient-physician relationship increases the revisit rate of the patient, or reduces arrival rate and queue length, hence freeing up additional server capacity. Moreover, we propose a partition of patients into a dual queue configuration system based on the patient level factors we combine to calculate a "benefit score", with higher benefit score patients being offered a dedicated queue service, and lower benefit score patients, a pooled queue.

2.2.2 Medical literature on care continuity

The medical literature differentiates between different types of care continuity, with Haggerty et al. (2003) providing a comprehensive review of the various forms, e.g., relational continuity, management continuity, and informational continuity. Relational continuity is the most studied of these – in fact, the terms relational continuity and continuity of care are often used synonymously in the existing literature – and is typically defined as the ongoing therapeutic

relationship between a patient and one or more providers. In this paper, the focus is on this relational component of continuity, as captured by repeat consultations with the same primary care physician.

It is well documented that continuity of care in primary care is valued by patients and doctors alike, with surveys of the extant medical literature highlighting various benefits of providing care continuity (The King's Fund 2016). In terms of direct benefits to patients, the medical literature has demonstrated various health benefits and improved management of health conditions for those who receive care continuity. For instance, studies have shown improvements in quality of life outcomes (Drury et al. 2020, Chen et al. 2017, Ye et al. 2016), blood pressure for diabetic and hypertensive patients (Leniz and Gulliford 2019), mortality (Maarsingh et al. 2016, Cho et al. 2015), adherence to medication plans (Dossa et al. 2017), and the likelihood of filling risky prescriptions (Hallvik et al. 2018).

In terms of indirect benefits, a meta-analysis by Huntley et al. (2014), involving participants from OECD countries, found that unscheduled secondary care usage is highly influenced by care continuity in the primary care setting. For example, primary care continuity has been associated with reductions in emergency department visits (Pourat et al. 2015) and unplanned hospitalizations of patients with ambulatory care sensitive conditions (Barker et al. 2017). Such advantages have been consistently demonstrated across different patient populations, including patients with serious mental illness (Ride et al. 2019), dementia (Amjad et al. 2016), COPD (Lin et al. 2015), and diabetes (Worrall and Knight 2011, Dossa et al. 2017), as well as older patients (Tammes et al. 2017, Katz et al. 2015, Bayliss et al. 2015, Nyweide et al. 2013). We contribute to this stream of literature by demonstrating that care continuity also affects the need for primary care visits themselves and that this effect is particularly pronounced in older patients and patients with complex conditions, such as chronic diseases or mental illnesses.

In summary, there is rich evidence to show the benefits of relational continuity for both patients and the wider health system in terms of reduced utilization. It is therefore somewhat surprising that the effect of continuity of care on clinical productivity within primary care practices has not yet been investigated. This study expands existing knowledge of the effects of care continuity by showing that care continuity not only improves outcomes and system utilization but also enhances the productivity of the primary care physicians themselves.

2.3 Hypothesis Development

Before we develop the paper's hypotheses, we clarify the key variables, care continuity and clinical productivity. The concept of relational continuity of care refers to a sustained therapeutic relationship between a patient and a doctor and is epitomized by the notion of a "patient list" or "patient panel" that many primary care physicians hold, either formally or informally (Wilkin and Metcalfe 1984, Tammes et al. 2017). These are the patients for whose health and care

the doctor takes personal responsibility over a prolonged period of time. Therefore, conditional on the physician being available, this study looks at care continuity as the the assignment of the patient to his preferred physician, and does not consider how the practice can change their scheduling policies to make the preferred physician easily available. The units of observation in this study are individual consultations, and we distinguish between appointments with a patient's regular doctor, who provides relational continuity and relational service, and appointments with another doctor providing one-off services, who we call a "transactional provider".

According to the Oxford Dictionary, productivity is defined as "the rate at which a worker, a company or a country produces goods, and the amount produced, compared with how much time, work and money is needed to produce them". It is therefore important to identify the locus of the productivity measurement – a worker, a company or a country. In this paper we study the worker – the individual GP – not the company – the practice. The difference is subtle but operationally important. Worker productivity is a measure of output per unit of time worked, while company productivity measures output per unit of worker time paid. There will be a difference because companies pay for productive as well as idle time. Company productivity is therefore affected by (i) worker productivity and (ii) the company's ability to keep the worker busy, e.g. by ensuring a steady flow of patients to a GP during the day. Addressing the latter scheduling question is outside the scope of this paper and our findings are therefore upper bounds for potential company productivity gains[1].

While the input measure for clinical productivity is clear – the clinician's time – defining an appropriate output measure is more difficult. On the one hand, the core activity of a primary care physician is providing consultations to patients. Accordingly, the average consultation duration or, equivalently, the number of patient consultations per clinician day is one relevant measure of productivity. On the other hand, such a daily throughput metric does not account for the recurrent nature of primary care services. A fast but ineffective patient consultation may create the need for more consultations in the future, while a longer but more thorough consultation may alleviate this need. Therefore, a slower doctor may actually be more productive overall. Consequently, to account for the recurrent nature of primary care services, we must consider two productivity metrics: (i) the duration of the visit, as a measure of throughput, and (ii) the revisit interval, i.e., the time between patient visits, as a measure of productivity-relevant service quality.

We note that in some contexts it might be highly salient to explore the relationship between visit

[1]In order to explore the potential gap between worker and company productivity in our context, we use a practice-level monthly panel to check the relationship between utilization (defined as number of patients per doctor session) and the continuity of care level at the practice. Surprisingly, we find a small positive correlation (0.05***), suggesting that higher continuity is associated with higher GP utilization within the context of current scheduling practices. A common practice is the use of a rota of "duty doctors" who provide urgent consultations on their duty days, instead of continuity of care consultations for their patients, and who often triage patients on the phone to decide whether they are better seen immediately by them or can wait for an appointment with their regular doctor.

duration and continuity of care. This relationship is determined by two countervailing factors. On the one hand, a doctor has an incentive to take more time to treat her regular patients thoroughly: this will reduce her future workload by preventing revisits, which would likely be her responsibility (Jeffers and Baker 2016). On the other hand, she is more familiar with these patients than a transactional provider and can therefore be more economical in her collection of information (Hill and Freeman 2011, Rosen et al. 2020).

However, given that this study uses UK data, we do not expect much difference in the duration of patients' consultations with their regular doctor and those with a transactional provider. This is because although it is not a formally imposed guideline, 10 minutes has become the de facto standard for primary care consultations in the UK (Royal College of General Practitioners 2019). For this reason, while we do provide exploratory evidence on the relationship between care continuity and consultation duration in the data in Section 2.6.4, we focus the hypothesis development on the relationship between care continuity and the revisit interval.

The idealised experiment is therefore to randomly assign either the patient's regular doctor or a randomly chosen other doctor from the same practice to each consultation and to measure operational productivity (duration of the consultation) and improved health. We use time to the next consultation to measure improved health.

2.3.1 Hypotheses

From a productivity standpoint, there are three critical differences between a patient's regular doctor and a transactional provider. First, a transactional provider shares the additional workload of a potentially avoidable revisit with all other doctors in the practice, while the patient's regular doctor will likely bear this future workload directly. As previously mentioned, this provides a stronger incentive for the regular doctor to provide service in a way that reduces the likelihood of a revisit and therefore her expected future workload. Second, the patient's regular doctor has better information about the patient than a transactional provider and may therefore be able to provide more effective customized advice and treatment that reduces the need for a revisit in the short term (Hjortdahl and Borchgrevink 1991). Moreover, information sharing leads to higher trust and trustworthiness, and consequently a more effective service provision (Özer et al. 2018). Third, the regular doctor has typically established a trust-based relationship with her patient, which provides rapport advantages and enables her to influence the patient more effectively (Tarrant et al. 2010, Hill and Freeman 2011). If the consultation is performed by the regular doctor, the incentive differential, information benefits, and relationship advantages interact in several ways to prolong the time to the next visit.

Specifically, the stronger incentive to reduce the likelihood of follow-up consultations makes the regular doctor more likely to diagnose the patient's problem carefully in an attempt to "get it right the first time" (Koopman et al. 2003). From a time and productivity perspective, such a root-cause diagnosis may not even cost the doctor very much if she is thoroughly familiar

with the patient. By contrast, a transactional provider has no particular incentive to go beyond the basic alleviation of a patient's presenting symptoms, knowing that any follow-up work is likely to be performed by a different doctor – an effect sometimes referred to as the "collusion of anonymity" (Freeman et al. 2010, Balint 1955). Furthermore, if a diagnosis is necessary, the transactional provider, being less familiar with the patient, is likely to need more time to arrive at a careful diagnosis. Given the time pressure of a full appointments schedule, a transactional provider is therefore more likely to adopt a trial-and-error approach to diagnosis and treatment, which increases the likelihood of a return visit whereas a regular doctor may practice 'wait and see' management of non-specific symptoms and consequently reduce undesirable medicalisations (Bobroske et al. 2021, Hjortdahl and Borchgrevink 1991).

Hypothesis 1. (H1) If a patient is seen by her regular doctor, she will have a longer time interval to her next appointment

A patient's regular doctor does not just have a stronger incentive to "get it right the first time" – they also have an incentive to leverage any patient encounter to explore the patient's health needs beyond the problem immediately at hand, as this may prevent an unnecessary visit in the near future (Hill and Freeman 2011). Thus, a regular doctor may check her notes and proactively deal with multiple illnesses or health issues in a single appointment as well as conduct opportunistic screening to detect problems early (Evans et al. 2008) In contrast, the conflict of responsibility for clinicians or 'collusion of anonymity' may occur when the patient sees a transactional provider who does not feel accountable for the patient (Balint 1955, Freeman et al. 2010). Transactional providers, lack both the incentive to expand their scope of service beyond the immediate clinical need expressed by the patient and the holistic patient knowledge that facilitates proactive management of the patient's health (Balint 1955). Existing literature has explored the relationship between tie strength and positive outcomes in such dyadic interactions (Sosa 2011). Opportunities for such proactive interventions, outside the scope of the immediate reason for the visit, are particularly salient when the patient has a chronic disease (Koopman et al. 2003, Goodwin et al. 2010). We therefore expect the productivity benefit of seeing a regular doctor to be larger for patients with multiple chronic conditions.

A regular doctor forms an impression of a patient's health over time and will therefore respond not only to the patient's current health status but also to *changes* in their overall health (Koopman et al. 2003). This longitudinal impression constitutes important additional information that a one-off transactional provider does not have (O'Connor et al. 1998, Ramanayake and Basnayake 2018). Indeed, the primary care physicians we interviewed for this study confirmed that when one of their regular patients enters the consultation room, they are likely to know immediately whether or not the patient is seriously ill and that they cannot do this with one-off patients. We expect that this ability to consider change in addition to state is particularly relevant for older patients, where single impressions of patient health can vary widely across patients and more information is gained from an observed change in an individual's health status. Being able to identify these changes and act early should allow the patient's regular doctor to reduce

the need for future visits.

Through repeated interactions over time, the regular doctor not only gains a comprehensive understanding of the patient's needs but also forms a trust-based relationship that may help her communicate more effectively with the patient (Tarrant et al. 2010, Özer et al. 2014). The patient may feel more secure sharing information with her regular doctor and, as a consequence, the doctor will be able to design a more appropriate treatment plan (von Bültzingslöwen et al. 2006). Another advantage of this rapport is that patients are more likely to comply with the doctor's instructions, making the treatment plan more effective and helping to prevent follow-up visits (Dossa et al. 2017, Brookhart et al. 2007). We expect the rapport advantage to be particularly important for patients with mental health conditions, such as anxiety, depression and schizophrenia (Biringer et al. 2017). The stigma that is still associated with mental health concerns might make patients reluctant to seek care and be fully transparent with an unfamiliar physician (Knaak et al. 2017). Moreover, medication compliance is a particular problem for mental health patients, due in part to the negative side effects associated with commonly prescribed medications (e.g. weight gain, fatigue, and reduced libido) (Semahegn et al. 2018).

Hypothesis 2. (H2) H1 will be larger for (a)patients with multiple comorbidities, (b)older patients and (c)patients with mental health conditions.

2.4 Clinical Setting, Data and Variables

In this section, we first provide a brief overview of the specifics of the UK primary care context that are relevant for this study. We then describe the dataset in detail and conclude with the description of the dependent variables, independent variables, and controls to be included in the analysis.

2.4.1 Primary care context

Although the English National Health Service (NHS) is publicly funded through taxation, primary care practices are privately owned small or medium-sized businesses, typically organized as partnerships of primary care physicians. Unlike hospitals, they are independent contractors with the NHS and therefore not under its direct control. Instead, the NHS controls their services through standard contractual arrangements. A typical practice has 8,000-10,000 registered patients, 4-5 full-time equivalent physicians, and a small number of other healthcare workers and administrative staff. Practice income is largely capitation-based, adjusted for demographic and socioeconomic characteristics of the registered practice population and geography. In 2019, a typical practice with 9,000 patients received an income of £1.3M, approximately £155 per registered patient or £32 per consultation (NHS Digital 2015).

The contract of a general practice in England stipulates a geographical catchment area for the practice. Patients who live in this area have the right to register with that practice. Patients may apply for registration with any practice but practices have discretion to accept or deny out-of-area patients. Importantly for this study, patients can only register with a single practice in England and are automatically deregistered when they register with a new practice. Since our data is practice-based, we therefore have full visibility of all primary care appointments of patients for their period of registration in the study practices.

Primary care services are free at the point of care. Patients can request to see any doctor at their practice, and practices generally try to ensure that a patient can see a preferred doctor. The NHS contract requires that each patient registered at a practice is assigned to a "named doctor," who is responsible for ensuring that the patient's needs are met (Tammes et al. 2017). However, practices may regard this as a purely administrative requirement, so the patient's "named doctor" may not be her regular doctor. This study focuses on the patient's regular doctor, the physician who has seen the patient most frequently in the past.

Consultations can be face-to-face, over the phone or video link or, in rare cases, at the patient's home. Appointments are generally booked via the phone with a receptionist. Practices have to accommodate routine and urgent appointments, the latter requiring on-the-day access. Some practices reserve a number of appointment slots for urgent services and, if those slots are booked, refer patients to the emergency department or ask them to call again the next day. Other practices accept all patients who call in before a certain cut-off time or offer unlimited access for acute care throughout the day. These practices have "duty doctors" dedicated to serving urgent care patients, typically in a round-robin fashion. Practices may also operate a phone-based triage system, with a doctor ascertaining whether a patient needs same-day service. While our data does not include the time when an appointment was made and therefore does not allow us to accurately distinguish between urgent and routine appointments, we are able to use markers, such as antibiotic prescriptions, that are more commonly associated with urgent appointments to help us distinguish the two appointment types.

2.4.2 Data and sample

In order to understand the effect of a patient seeing her regular doctor in one consultation on the time to her subsequent consultation, we perform a cross-sectional analysis with individual patient consultations as observations. We obtained consultation-level data from the UK Clinical Practice Research Datalink (CPRD). This database consists of anonymized electronic medical records covering over 11.3 million patients (6.9% of the population of the UK) across 674 practices in the UK; it is representative of the population in terms of age, sex and ethnicity (Herrett et al. 2015). The database encompasses a wealth of information about patients, visits, providers, diagnoses, prescriptions, referrals, treatments, immunization records, and test records. It can be linked to several other data sources such as secondary care services, enabling a fairly complete

Criteria	Patients	Consultations
Doctor consultations in 407 primary care practices	5,475,342	370,890,526
Face-to-face doctor consultations only	5,335,945	200,312,789
Consultations after date at which practice data is of research quality	5,037,650	161,556,335
Consultations during a patient's continuous registration period	4,921,208	139,455,412
Consultations at which the patient was over 18	3,855,445	86,399,813
Consultations with ≥ 3 and ≤ 104 consultations in the preceding 2 years	2,952,445	71,797,380
Consultations occurring after the patient's first two years following registration	2,537,781	63,087,124
Only consultations with a valid revisit interval	2,410,189	60,894,300
Only consultations between January 2007 and December 2017	2,322,773	51,711,037
Consultations occurring when the patient's regular doctor is available	2,273,571	45,376,070
Random sample of consultations from 381 remaining practices	1,883,626	11,344,065

Table 2.1 Data and sample inclusion criteria

medical record to be obtained for a given patient.

We obtained data for all English practices that had consented to linkage to secondary care usage data. The restriction to English practices improves the homogeneity of the sample as the national health systems operate differently in the four constituent countries of the UK and there are differences in their standard primary care contracts. The starting data set comprised information on 370,890,526 primary care consultations corresponding to 5,475,342 patients. The analysis sample was derived from this data using the inclusion criteria described in Section 3.2 of Chapter 3 and summarized in Table 2.1.

Specifically, the sample for analysis consists of face-to-face consultations between primary care doctors and patients over the age of 18 that took place between January 2007 and December 2017 at times during which (i) a practice's data is deemed to be of research quality, (ii) the patient had sufficient past appointments to identify their regular doctor, and (iii) the patient's regular doctor was available to see them. Data was also filtered to exclude patients visiting more than once per week on average, to only include data from periods in which the patient was continuously registered at a practice, and to exclude any consultations lacking an observable revisit interval. (For more justification on the inclusion criteria, see Section 3.2 of Chapter 3.)

The final sample thus consists of 11,344,065 consultations between 1,883,626 patients and 14,123 doctors in 381 practices across 11 years between 2007 and 2017.

2.4.3 Variable description

2.4.3.1 Dependent variable.

The main dependent variable throughout this study is the patient's revisit interval (RI), which is defined by the time elapsed between the focal face-to-face consultation with a doctor and the next face-to-face consultation with a doctor, and is measured in days. The hypotheses posit

that for a consultation with the patient's regular doctor, the revisit interval will be longer than it is for a consultation with a transactional provider. As is usually the case with durations, the distribution of revisit intervals is right-skewed. We therefore transform the variable to a logarithmic scale by taking it's natural logarithm. Figure 2.1 shows the distribution of the log-transformed revisit interval, and Table 2.2 contains summary statistics. In the infrequent event that a patient has multiple face-to-face consultations with doctors on the same day, we set the revisit interval length to 0.5 days prior to log transforming.

We believe that pure throughput metrics, such as length of stay or number of appointments per day, do not capture productivity in a context with repeated interactions. Therefore, we focus on prolonging the revisit interval, which is likely to signal both productivity and quality benefits.

Since we consider a non-traditional quality and productivity measure, we wish to ensure that other quality outcomes do not change, if not become even better. Therefore, as additional checks we also consider alternative dependent variables of quality such as referral rate, prescription rate, emergency department admissions (Section 3.15 of Chapter 3).

2.4.3.2 Independent variables.

The main independent variable is a binary variable, which indicates whether or not the focal consultation was with the patient's regular doctor.

When the patient calls in for an appointment, the patient can request to see any doctor at their own practice. Generally, patients have a preference for a certain doctor who they have a good relationship with. For routine appointments, the norm is that the practice generally tries to accommodate the request of the patient to see their preferred doctor. Doctors also hold informal 'lists' of patients who they have a regular relationship with and they feel responsible and accountable foe their patient list. Practices also have to accommodate urgent appointments that require on-the-day access. In most cases, practices have certain doctors assigned to deal with urgent requests and are unable to schedule the patient with their preferred provider. Patients with non-acute conditions may be willing to wait longer to see their preferred doctor and might not need to return for a follow-up soon, or patients with acute conditions might be willing to accept an appointment with any doctor and also return soon for a follow-up appointment, creating an endogeneity issue that we address in our empirical strategy. Since we are unable to observe in our data when the appointment was booked, we use proxy markers to help us distinguish the types of appointments.

Since this study spans an 11-year time horizon, the patient's regular doctor may change over time for various reasons. For example, a doctor may retire or leave the practice, or the patient may switch due to a change in circumstances or a positive experience with another doctor. When we determine a patient's regular doctor we therefore use a dynamic measure rather than identifying one fixed dyad for each patient.

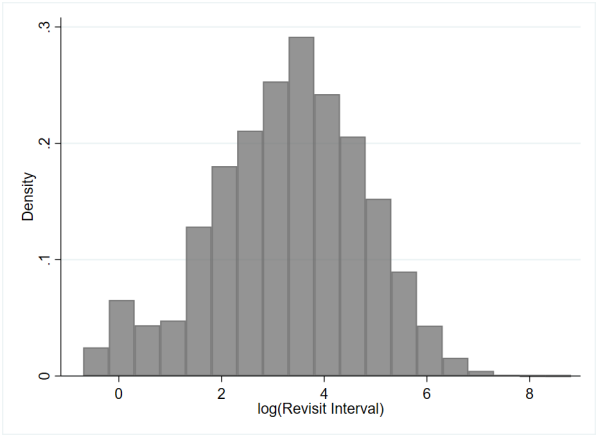

Figure 2.1 Distribution of ln(Revisit Interval)

	ln(Revisit Interval)
Mean	3.29
Median	3.33
Min	-0.69
Max	8.34
Std. Dev.	1.48

Table 2.2 Descriptive statistics for the dependent variable

Specifically, we consider a two-year time window over which we calculate the patient's regular doctor. For consultation i we define a patient's regular doctor as the doctor with whom the patient had the most face-to-face consultations over the preceding two year. To break ties, we choose established over unestablished doctors[2] and, if the tie persists, we choose the doctor who the patient saw most recently. The independent variable RD_i is thus a binary variable that is equal to 0 for consultation i when the patient does not see her regular doctor and 1 for consultation i if the patient does see her regular doctor. Overall, 50% of consultations in our analysis sample occur between a patient and their regular doctor. We provide additional information on the independent variable and the patient-level factors that affect it in Section 3.3 of Chapter 3.

Since we update the independent variable at every consultation, the measure also accounts for the fact that patients might want to see a different doctor due to a positive experience with another doctor. Moreover, in our analysis we account for the degree of the tie between the patient and the doctor, for example, whether the patient sees the regular doctor 2 out of 10 times or 8 out of 10 times (Section 3.17 of Chapter 3).

We believe that our measure captures the notion of the patient's regular doctor's accountability and responsibility for the health of the patient and hence is most appropriate for our setting. There are also other measures of continuity of care in the literature, which we describe in Section 3.4 of Chapter 3.

2.4.3.3 Clinical productivity and consultation-level analysis

Using the dependent and independent variable as described, we look at whether conditional on the doctor being available, seeing the regular doctor has the potential to significantly increase

[2]The data allow us to distinguish between two types of physicians, *established* physicians who have a contract with the practice and *unestablished* physicians who are not permanent employees of the practice but may be self-employed or employed with an agency. The latter work on an ad-hoc basis, often in multiple practices, and are paid on an hourly basis.

clinical productivity by 'sorting patients out' for longer. Therefore, from a clinical productivity perspective, this translates to getting more out of the physician's time. We define clinical productivity as:

$$\frac{ImprovedHealth}{ClinicianHour} = \frac{ImprovedHealth}{Consultation} * \frac{Consultation}{ClinicalHour}$$

In other words, clinical productivity is the product of quality and throughput, which makes a consultation-level analysis natural. We use time to the next consultation to measure improved health and the duration of the consultation to measure throughput.

The study assumes that the practice environment constant and does not consider what happens at the population level if a practice increases continuity of care at the expense of rapid access. In this paper, we do not consider how the practice can make the regular doctor more available, as this is a different management questions that warrants a separate analysis.

2.4.3.4 Control variables.

An important confounding factor that need to be accounted for is the patient's visit history. The more frequently a patient visited in the past, the more likely they are to also visit frequently in the future. Hence, we expect that such a patient will have a shorter revisit interval. At the same time, a patient's likelihood of a consultation with a regular doctor is affected by her consultation frequency (see Figure 3.1). Since we anticipate nonlinear effects (see Figure 3.1), we include both the patient's past consultation frequency as a categorical control as well as her average past revisit interval as a linear continuous control (see Table 2.3 for details). The categories are shown in Figure 3.6 of Chapter 3.

The relationship between continuity of care and a patient's revisit interval is also likely to be confounded by patient demographic factors, attributes of the regular doctor, temporal factors, and practice-level factors. To account for these potential confounds, we use several control variables. A summary of the controls is provided in Table 2.3 and an explanation for each of the controls is given in Section 3.6 of the Chapter 3.

Table 2.3 Table of controls

Variable	Type	Description
Patient demographics		
Age	Categorical (14)	Age of the patient at the time of consultation, split into age bands (18-25, 26-30, 31-35, . . ., 81-85, 86+)
Number of comorbidities	Categorical (6)	Number of comorbidities at the time of consultation, calculated using the Cambridge Comorbidity Index (CCI), split into bands (0, 1, . . ., 4, 5+)
Individual comorbidities	Binary (26)	For each of the 26 comorbidities as defined by the CCI, a variable to indicate whether the patient suffers from that comorbidity at the time of the consultation
Mental health	Binary	A variable to indicate whether the patient suffers from a mental health condition as defined by the CCI
Gender	Binary	A variable which is 0 if the patient is female and 1 if the patient is male
Index of Multiple Deprivation	Categorical (5)	The deprivation level assigned to the patient
Prescriptions	Categorical (10)	The number of repeat prescriptions the patient is prescribed within the 6 months preceding the focal consultation, split into bands (0, 1, 2, 3, 4-5, 6-7, 8-9, 10-12, 13-15, 16+)
Patient's past history		
Past consultation frequency	Categorical (20)	The total number of consultations in the 2 years preceding the patient's focal consultation, split into 20 bands: 3-10 consultations (8 categories of size 1), 11-20 consultations (5 categories of size 2), 21-26 consultations (2 categories of size 3), 27-30 consultations (1 category of size 4), 31-35 consultations (1 category of size 5), 36-55 consultations (2 categories of size 10), 56+ consultations (1 category).
Past revisit interval	Continuous	The 2-year past average revisit interval of the patient calculated as described in Section 3.8.3 of the Chapter 3.
Attributes of the regular doctor		
Established doctor	Binary	A variable to indicate if the patient's assigned regular doctor for the focal consultation is an established or unestablished doctor (see Footnote 2)
Temporal factors		
Year	Categorical (11)	Year during which the consultation took place (2007-2017)
Month of Year	Categorical (12)	Month of the year in which the consultation falls (January-December)
Day of Week	Categorical (7)	Day of the week in which the consultation took place (Monday-Sunday)
Practice-level factors		
Practice-level demand	Continuous	Total practice demand during the focal week of each consultation, standardized by a weekly average in a 52-week period around the focal week
Practice	Categorical (381)	The practice at which the consultation took place

Notes: If a variable is categorical, the number in (·) in the " Type" column indicates the number of levels.

2.5 Econometric Specifications

2.5.1 Ordinary least squares estimator

To identify the effect of a patient's seeing her regular doctor on her revisit interval (Hypothesis 1 (H1) in a robust way, we estimate a series of consultation-level models using the revisit interval RI_i as the dependent variable and the indicator RD_i for the regular doctor as the main explanatory variable, where the index i refers to a consultation. We begin with a standard OLS model:

$$ln(RI_i) \quad = \quad \beta_0 + \beta_1 RD_i + \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i \ , \tag{2.1}$$

where the vector $\mathbf{X}_i$ specifies the set of controls corresponding to consultation i (as defined in Table 2.3) and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is the error term. We cluster standard errors at the patient level to account for correlations of error terms when consultations are associated with the same patient. The effect of interest is captured by the coefficient β_1, where H1 posits that $\beta_1 > 0$.

2.5.2 Acuity subsamples

Given the observational nature of the data, we are naturally concerned about confounding. Patient acuity is an obvious example of an omitted variable that may bias the OLS results: patients with higher acuity may be unable to wait for an appointment with their regular doctor, so they may be more likely to see a non-regular doctor. At the same time, because of the acute nature of their condition, these patients may also require near-term follow-up appointments, leading to a shorter revisit interval. This could provide an alternative, non-causal explanation for a positive β_1 in (2.1). As we cannot directly measure patient acuity, nor do we observe when the appointment was booked, adding control variables to account for heterogeneity in acuity is not feasible in this case.

To assess the potential impact of acuity on the results, we use two subsamples of consultations with characteristics that are more likely in acute presentations: (i) the subsample of consultations where the patient was prescribed an antibiotic, and (ii) the subsample of consultations with patients who visited an emergency department (ED) in the seven-day window prior to the focal consultation. Both an antibiotic prescription and a prior ED visit are indicative of a consultation for an acute condition. Hence, if acuity confounds the effect, we would expect a substantially smaller and perhaps insignificant coefficient β_1 in (2.1) when estimated on these subsamples.

2.5.3 Instrumental variable estimators

To address general endogeneity concerns, we use two instrumental variable specifications, control functions and two-stage least squares.

2.5.3.1 Control functions.

The control function (CF) approach is based on estimating the following selection and outcome equations

$$RD_i^* = \alpha_0 + \mathbf{X}_i\boldsymbol{\alpha} + \alpha_{n+1}IV_i + \delta_i\,, \;\; RD_i = \mathbb{1}[RD_i^* > 0]\,, \tag{2.2}$$

$$ln(RI_i) = \beta_0^{CF} + \beta_1^{CF}RD_i + \mathbf{X}_i\boldsymbol{\beta}^{CF} + \gamma_1^{CF}\widehat{RD_i} + \epsilon_i^{CF}\,, \tag{2.3}$$

where RD_i^* is a latent variable, $\delta_i \sim \mathcal{N}(0,1)$, $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$, $\mathbb{1}[\cdot]$ is the indicator function, $\mathbf{X}_i$ is the $N \times n$ matrix of controls, IV_i is an instrumental variable (to be described shortly), and $\widehat{RD_i}$ is the generalized probit residual of observation i.

Following the probit estimation of Equation (2.2) as a first stage, $\widehat{RD_i}$ is calculated as

$$\widehat{RD_i} = \frac{\phi(\mathbf{X}_i'\boldsymbol{\alpha}')\,[RD_i - \Phi(\mathbf{X}_i'\boldsymbol{\alpha}')]}{\Phi(\mathbf{X}_i'\boldsymbol{\alpha}')\,[1 - \Phi(\mathbf{X}_i'\boldsymbol{\alpha}')]}\,,$$

where $\mathbf{X}_i'\boldsymbol{\alpha}' = \alpha_0 + \mathbf{X}_i\boldsymbol{\alpha} + \alpha_{n+1}IV_i$, and $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and cumulative distribution functions of the standard normal distribution, respectively. The estimation is then completed by estimating the second stage given in Equation (2.3).

The CF approach is similar to the more common two-stage least square (2SLS) method. However, in contrast to 2SLS, the CF approach estimates a probit model (i.e., Equation Equation 2.2) in the first stage and then uses the generalized probit residual $\widehat{RD_i}$ as an additional control in the outcome equation (i.e., Equation Equation 2.3). The addition of the generalized probit residual then adjusts the coefficient β_1^{CF} for unobserved confounders that affect both the endogenous regressor RD_i and the dependent variable $ln(RI_i)$ (Wooldridge 2015). The t-statistic of the coefficient γ^{CF} can be used in a straightforward manner to test for endogeneity in the CF model (Wooldridge 2002).[3]

Consistency of the CF approach requires the probit model to be a correct specification for the likelihood of seeing the regular doctor, i.e., $P(RD_i = 1|\mathbf{X}_i') = \phi(\mathbf{X}_i'\boldsymbol{\alpha}')$. In contrast, the 2SLS estimator does not impose strict distributional assumptions on $P(RD_i = 1|\mathbf{X}_i')$. However, using the standard 2SLS estimator with a nonlinear model in the first stage renders the estimates inconsistent (Wooldridge (2002) refers to this as the "forbidden regression"). An alternative is a 2SLS approach with a linear probability specification in the first stage, i.e., replacing the probit first stage with an OLS estimate. We use this method as a second IV approach to corroborate the CF estimates.

2.5.3.2 The instrumental variable.

Both the CF and 2SLS approaches rely on the availability of an instrumental variable (IV) (Wooldridge 2002). The IV should affect the probability that the patient will see her regular

[3] We estimate the CF model using Stata's etregress command with the two-step option and bootstrapped standard errors. We refer to Wooldridge (2002) and Wooldridge (2015) for a technically detailed explanation of the CF approach.

doctor, so it should be significant in the first stage equation (i.e., it should be relevant), but it should not affect the dependent variable $ln(RI_i)$ except through the independent variable RD_i (i.e., it should be valid). In this study, we use an IV that captures whether the focal patient's regular doctor is relatively more accessible for her regular patients during the week of the focal consultation, compared to her long-run average accessibility. This measure is calculated using the set of patients who share the same regular doctor as the focal patient at the time of the consultation, but it excludes any visits by the focal patient herself. The formal description of the IV calculation is given in Section 3.8 of Chapter 3.

We believe this is a relevant IV in our study context. A doctor who is more (or less) accessible than usual to her other regular patients is also likely to be more (or less) accessible to the focal patient. Consistent with this intuition, we find a fairly large correlation between the IV and the patient's seeing her regular doctor ($\rho = 0.16$, $p < 0.001$), indicating that the IV is likely both relevant and strong. Formal hypothesis testing for under- and weak identification, reported in Section 3.8.1 of Chapter 3, provide strong evidence that the instrument is relevant and the endogenous regressor is not weakly identified.

Turning to the validity condition, there is no reason to believe that the accessibility of the regular doctor for other patients should directly affect the revisit interval of the focal patient. Yet, it is possible that there are unobserved factors that correlate with both the relative accessibility of the regular doctor for other patients and the focal patient's revisit interval (e.g., a flu outbreak that reduces the expected revisit interval of the focal patient and also makes it harder than normal for patients to access their regular doctors). Importantly, however, such factors should affect the revisit interval not only of the focal patient but also of the doctor's other patients. Therefore, it is possible to account for these unobservable factors by adding as a control variable the average of $ln(RI)$ of other patients who (i) share the same regular doctor as the focal patient and (ii) visit a doctor in the same week as the focal patient. When the average revisit interval of other patients changes (e.g., due to a flu outbreak), then this control variable also adjusts the expected revisit interval of the focal patient in the same direction. This control variable thus helps to account for unobserved factors correlated with both the IV and the outcome, thus strengthening the validity of the IV. A similar approach is used in Freeman et al. (2020) and Bobroske et al. (2021). The full description of this control is provided in Section 3.8.3 of Chapter 3.

2.5.4 Propensity score-based estimation.

To validate the IV approaches, we estimate a propensity score-based matching estimator known as the minimum bias estimator (MBE). Matching itself can alleviate the impact of unobserved bias and, moreover, the MBE is an approach designed to address endogeneity bias without relying on an instrumental variable. The MBE thus provides an alternative method to corroborate the results (Rosenbaum 2005).

2.5.4.1 Constructing the matched sample.

To implement the MBE approach, we first reduce the sample to include only one randomly chosen consultation per patient. This ensures that a patient is not matched with themselves when applying the matching procedure to the data. However, this reduced sample is not representative of the original sample. Every patient is represented only once, independently of the frequency of their use of primary care, and therefore the reduced sample is biased towards consultations with healthier and younger individuals relative to the original sample of consultations. We address this bias by drawing a sample of 25% of the reduced consultation sample using the frequencies of the patients' visits over the entire observation period as probability weights (Jann 2006). This probability weighting biases this second sample towards more frequent users of primary care (i.e., the older and less healthy population) and produces a more representative subsample of consultations.

Using this subsample of consultations, we next match consultations in the control group (those with a transactional provider for the patient's randomly chosen consultation) to the treated group (those with the patient's regular doctor for their randomly chosen consultation) using nearest neighbor matching without replacement. Matching is performed on propensity scores generated from a probit regression that is estimated based on the full set of covariates included within the control matrix $\mathbf{X}_i$ in Equation Equation 2.1. We also include the condition that the maximum distance between the propensity scores of two observations chosen as potential neighbors (i.e., the caliper) is 0.001. This narrow caliper helps reduce the potential for bias when examining the difference between the average revisit intervals in the two subsamples.

The matched sample consists of 494,810 consultations, half of whom saw their regular doctor and half who did not. Summary statistics comparing the covariate profile of the full analysis sample and the matched sample are provided in Section 3.10 of the Chapter 3.

2.5.4.2 Average treatment effect estimation.

Using the matched sample, we estimate the average treatment effect (ATE) in two ways. First, we report the difference in averages of $ln(RI)$ between the control and the treated groups. Second, we re-estimate the OLS regression specified in Equation Equation 2.1 using the matched sample in order to control for the effect of other covariates. Generally, regression after matching is not recommended because standard errors in the regression do not correct for the fact that the matching step has already been performed. However, Austin and Small (2014) and Abadie and Spiess (2021) show that this can be addressed by correcting standard errors by clustering them at the level of the matched pair (i.e., two observations per cluster). We follow this recommendation when reporting effects.

2.5.4.3 Minimum bias estimation.

We combine the treatment effect estimations for the matched samples with the minimum bias estimator (MBE), specified by Millimet and Tchernis (2013). This method aims to minimize the effect of unobserved bias and the impact of omitted variables without relying on instrumental variables and exclusion restrictions. The approach is motivated by the observation that any hidden bias has the biggest impact on the tails of the distribution of selection probabilities (i.e., propensity scores closer to 0 or 1, as calculated using the probit model described in Section 2.5.4.1). It then follows that the bias is minimized for matched observations with propensity scores equal to 0.5 (the score which is closest to the probability of random assignment) (Peel 2018). The impact of the potential hidden bias is therefore minimized by restricting the sample of matched cases to those with propensity scores within a defined narrow interval around 0.5 (Peel 2018). The direction and strength of the bias can be assessed by widening that interval. Practically, this means that ATEs are estimated on subsamples that are formed by narrowing the range of propensity scores. Black and Smith (2004) recommend a range between 0.33 and 0.67, though with more data (as in our case), this interval can be restricted further.

2.6 Results

Table 2.4 shows the OLS estimates, based on Equation Equation 2.1, for various control structures. Controls are added step-wise into the model to help us understand how the inclusion of different control categories alters the main effect estimate. The step-wise introduction of factors that are correlated with poorer heath, such as consultation frequency, past average revisit interval, comorbidity controls and age and socioeconomic factors, increase the effect of seeing the regular doctor. This is suggestive of heterogeneity, which we will address directly in Section 2.6.3.

Table 2.4 Sensitivity of OLS coefficient estimates to the inclusion of different categories of controls

	OLS: Dependent variable = Natural logarithm of the revisit interval				
$RD_i = 1$	0.050^{***} [0.048,0.052]	0.112^{***} [0.110,0.114]	0.150^{***} [0.149,0.152]	0.150^{***} [0.148,0.152]	0.150^{***} [0.148,0.151]
Patient demographics	No	Yes	Yes	Yes	Yes
Patient's past history	No	No	Yes	Yes	Yes
Attributes of regular doctor	No	No	No	Yes	Yes
Practice-level demand	No	No	No	No	Yes
Temporal factors	Yes	Yes	Yes	Yes	Yes
Practice FE	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.045	0.132	0.153	0.158	0.160
Number of observations	11,344,065	11,344,065	11,344,065	11,344,065	11,344,065

Notes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$; 95% confidence intervals in square brackets, with standard errors clustered at the patient level; Variables included within the categories of controls are specified in Table 2.3.

Examining the results, we find evidence that patients' consultations with their regular doctor are robustly associated with a longer revisit interval ($\beta_1 = 0.150$, 95% CI: $[0.148, 0.152]$, p-value < 0.001 in the fully controlled model). Since the dependent variable has a log-scale, the full model suggests that the revisit interval increases by 15.0%, on average, when a patient is seen by their regular doctor.

2.6.1 Acuity subsamples

Estimating the fully controlled OLS model on the subsample of consultations with antibiotic prescriptions provided a smaller coefficient $\beta_1 = 0.114$ (95% CI: $[0.110, 0.119]$), suggesting that the continuity of care effect may be less pronounced for acute conditions. This is expected, as after an acute visit, a regular doctor is more likely to see the patient again for a follow-up and may have less leeway to extend the revisit interval length. However, the coefficient still remains significantly positive and the effect remains large. This observation is corroborated by the second subsample analysis, consisting of consultations with patients who had visited an ED in the seven-day window prior to the focal consultation. Again, the coefficient is lower than in the full sample ($\beta_1 = 0.098$, 95% CI: $[0.084, 0.112]$). However, the effect remains statistically highly significant and the 9.8% estimated average extension of the revisit interval remains practically significant as well.

In summary, the subsample analyses suggest that confounding by acuity, if it occurs, is relatively small and does not explain the main effect. The analyses are fully documented in Section 3.9 of the Chapter 3.

2.6.2 Alternative model specifications

Table 2.5 summarizes the results from all models introduced in Section 2.5. The top panel reports the estimated effect size associated with seeing the regular doctor (RD_i) on the natural logarithm of the revisit interval using OLS (1a), the CF model (1b), and the 2SLS model (1c). The subsequent panels correspond to estimates using propensity score matching (PSM). All PSM coefficients are estimated using the reduced sample, where a single consultation was randomly chosen per patient, as explained in Section 2.5.4.1. For PSM (2a) we report differences in averages of the revisit interval, $ln(RI)$, between the control and treated groups; PSM OLS (2b) corresponds to an OLS regression on the matched sample that includes controls (see Section 2.5.4.2). We then repeat PSM and PSM OLS using three different propensity ranges for the MBE (3a-5b). Narrower propensity ranges correspond to a less biased estimate (see Section 2.5.4.3).

Table 2.5 Coefficients of RD_i (seeing the regular doctor) on $ln(RI_i)$ (log revisit interval) for different model specifications

Dependent variable = Natural logarithm of the revisit interval								
Model	Sample	Method	Coefficient (β_1)	Std. Error	t-statistic	P> $\|t\|$	95% CI	
1a	consultations	OLS	15.0%	0.1%	182.61	0.00	14.8%	15.2%
1b	consultations	CF	13.2%	0.4%	26.61	0.00	12.2%	14.1%
1c	consultations	2SLS	12.5%	0.5%	24.04	0.00	11.5%	13.6%
2a	patients	PSM	16.1%	0.4%	39.05	0.00	15.3%	16.9%
2b	patients	PSM OLS	16.1%	0.4%	42.19	0.00	15.4%	16.9%
3a	patients	PSM 0.25<p<0.75	16.1%	0.4%	37.91	0.00	15.3%	16.9%
3b	patients	PSM OLS 0.25<p<0.75	16.1%	0.4%	42.19	0.00	15.3%	16.9%
4a	patients	PSM 0.33<p<0.67	16.0%	0.5%	35.09	0.00	15.1%	16.9%
4b	patients	PSM OLS 0.33<p<0.67	16.1%	0.4%	38.23	0.00	15.3%	16.9%
5a	patients	PSM 0.4<p<0.6	16.6%	0.5%	30.70	0.00	15.5%	17.7%
5b	patients	PSM OLS 0.4<p<0.6	16.5%	0.5%	33.12	0.00	15.5%	17.5%

Notes: Standard errors clustered at the patient level for models 1a-1c and the matched pair level for models 2a-5b. This table reports the coefficient estimates of RD_i (seeing the regular doctor) for the taxonomy of models specified in Section 2.5. OLS regression refers to the ordinary least squares regression Equation 2.1. CF refers to the control function approach specified in Section 2.5.3.1 and 2SLS refers to the two-stage least squares method using the same IV as used for the CF approach. PSM corresponds to the matching-based effect estimation where we report differences in averages of $ln(RI)$ (log revisit interval) between the control and treated groups, using nearest neighbor matching without replacement and a caliper of 0.001 (Section 2.5.4.2). PSM OLS corresponds to an OLS regression on the matched sample that includes the covariates (Section 2.5.4.2). Models 3a-5b correspond to the minimum bias estimators that use subsamples determined by the propensities (p) (Section 2.5.4.3).

We find that the results are consistent across all modeling techniques employed and confirm H1. While the coefficient of the bias correction term in the CF model is statistically significant ($\hat{\gamma}^{CF} = 0.014$, 95% CI: $[0.009, 0.021]$), there is little evidence of major confounding, with all

estimates close to the 15% effect size estimated by the original OLS model. The MBE estimates do not change significantly as we narrow the propensity score ranges and the direction of the change is towards higher values of the coefficient for narrower ranges.

As the CF approach is both conservative and indicated for use in contexts with a binary first stage selection equation (see Section 2.5.3.1), we select this approach as the main model specification going forward. The CF method estimates that patients who see their regular provider have a 13.2% longer revisit interval (95% CI: $[12.2\%, 14.1\%]$).

2.6.3 Moderating effects

The effect of continuity of care is likely to be heterogeneous across patient segments. In this section, we test the moderating effects of comorbidity (H2a), age (H2b) and mental health (H2c). To do so, we re-estimate the CF model (Equations Equation 2.2 and Equation 2.3) but now also include interaction terms between the main independent variable RD_i and the three patient-level moderation variables. As the moderators are naturally correlated (in particular age and comorbidity), we have included all moderators in a single model to estimate their effect net of the correlated moderating effects of the other variables.

Estimates of the average marginal effects based on the moderation results are reported in Table 2.6 and a graphical representation of the results is given in Figure 3.8 in Chapter 3. Table 2.6 reports the estimated revisit interval for an average individual (within the segment specified by the first column of the row) assuming they either saw a transactional provider ($RD_i = 0$ columns) or their regular doctor ($RD_i = 1$ columns). For example, an average 18-25 year old who saw a transactional doctor is estimated to have a (natural logarithm of the) revisit interval of 3.31, as compared to 3.35 if they had instead seen their regular doctor. The difference between these two values is given by the effect size, e.g., 4.1% (95% CI: $[2.9\%, 5.3\%]$) for an average 18-25 year old patient. (The coefficients corresponding to the moderating effect estimates are reported in Table 3.11 in Chapter 3.)

Table 2.6 Average marginal effects associated with seeing the regular doctor when using age, comorbidity and mental health as moderators, calculated using the control function model

	$RD_i = 0$		$RD_i = 1$		Effect Size	
	ln(RI)	95% CI	ln(RI)	95% CI	Mfx.	95% CI
Comorbidity						
0 comorbidites	3.33	[3.32,3.34]	3.42	[3.42,3.43]	9.3%	[8.3%,10.3%]
1 comorbidity	3.24	[3.24,3.25]	3.36	[3.36,3.37]	12.2%	[11.1%,13.2%]
2 comorbidites	3.19	[3.19,3.20]	3.33	[3.32,3.33]	13.4%	[12.4%,14.4%]
3 comorbidites	3.17	[3.16,3.18]	3.30	[3.30,3.31]	13.5%	[12.4%,14.5%]
4 comorbidites	3.16	[3.16,3.17]	3.29	[3.28,3.30]	12.9%	[11.8%,14.0%]
≥ 5 comorbidites	3.15	[3.14,3.16]	3.27	[3.26,3.28]	11.9%	[10.8%,13.0%]
Age band						
18-25yrs	3.31	[3.30,3.31]	3.35	[3.34,3.35]	4.1%	[2.9%,5.3%]
26-30yrs	3.29	[3.29,3.30]	3.35	[3.34,3.35]	5.5%	[4.3%,6.7%]
31-35yrs	3.30	[3.29,3.30]	3.35	[3.35,3.36]	5.8%	[4.6%,6.9%]
36-40yrs	3.31	[3.30,3.31]	3.37	[3.37,3.40]	6.4%	[5.2%,7.5%]
41-15yrs	3.31	[3.30,3.32]	3.39	[3.38,3.40]	8.1%	[7.0%,9.2%]
46-50yrs	3.30	[3.29,3.30]	3.39	[3.39,3.40]	9.6%	[8.5%,10.7%]
51-55yrs	3.29	[3.29,3.30]	3.40	[3.39,3.41]	10.7%	[9.7%,11.8%]
56-60yrs	3.28	[3.27,3.28]	3.40	[3.39,3.41]	12.3%	[11.2%,13.4%]
61-65yrs	3.26	[3.25,3.27]	3.40	[3.39,3.40]	13.4%	[12.4%,14.5%]
66-70yrs	3.23	[3.23,3.24]	3.38	[3.37,3.38]	14.4%	[13.3%,15.5%]
71-75yrs	3.19	[3.18,3.20]	3.35	[3.34,3.36]	15.9%	[14.8%,16.9%]
76-80yrs	3.14	[3.14,3.15]	3.31	[3.31,3.32]	17.2%	[16.1%,18.3%]
81-85yrs	3.08	[3.07,3.09]	3.26	[3.25,3.27]	17.9%	[16.7%,19.0%]
86+yrs	2.97	[2.96,2.98]	3.14	[3.14,3.15]	17.6%	[16.5%,18.7%]
Mental Health						
No	3.23	[3.23,3.24]	3.35	[3.34,3.35]	11.2%	[10.2%,12.2%]
Yes	3.22	[3.21,3.23]	3.35	[3.35,3.36]	13.4%	[12.4%,14.4%]

Notes: '$RD_i = 0$' (resp., '$RD_i = 1$') columns specify the estimated natural logarithm of the revisit interval (ln(RI)) together with 95% confidence intervals (CI) for a patient who saw a transactional (resp., regular) provider; 'Effect Size' columns give the average marginal effect (mfx.), with 95% CIs, associated with a patient seeing a regular doctor versus a transactional provider, implemented using Stata's margins command.

2.6.3.1 Comorbidity.

The comorbidity panel in Table 2.6 confirms that patients who see their regular doctor ($RD_i = 1$ columns) have, on average, longer revisit intervals than those who see a transactional provider ($RD_i = 0$ columns), and this effect is independent of the number of comorbidities. The effect size columns confirms H2a by showing that the extension of the revisit interval that the regular doctor achieves is higher for patients with comorbidities. The difference between zero and one comorbidities is statistically significant. Additional differential effects beyond two comorbidities are insignificant.

2.6.3.2 Age.

The age band panel in Table 2.6 shows the effect of patient age. As in the case of comorbidity interactions, the table confirms that patients who see their regular doctor ($RD_i = 1$ columns) have longer revisit intervals than those who see a transactional provider ($RD_i = 0$ columns). The effect size columns shows the difference and confirms H2b: seeing a regular doctor is particularly productive for older patients. The extension of the revisit interval length increases from 4.1% (95% CI: $[2.9\%, 5.3\%]$) for 18-25 year-old patients to 17.6% (95% CI: $[16.5\%, 18.7\%]$) for patients over 86.

Further analysis of the interplay between age and comorbidity, reported in Section 3.13 of Chapter 3, shows that the moderating effect of comorbidities is most pronounced in younger patients, becomes weaker with age, and disappears for patients over the age of 80.

2.6.3.3 Mental health.

Since mental health patients are more likely to be heavy users of primary care, they have a slightly shorter revisit interval on average than patients without mental health conditions. However, though the difference is small, we find that providing continuity to such patients extends their revisit interval by 13.4% (95% CI: $[12.4\%, 13.4\%]$), compared to a 11.2% (95% CI: $[10.2\%, 12.2\%]$) improvement for patients without such conditions.

2.6.4 Duration of consultations

As mentioned in Section 2.3, in addition to the length of the revisit interval, which is the focus of our analysis, consultation length is a second important productivity measure. On the one hand, consultations between a patient and her regular doctor may take longer as the doctor may take more time to avoid a potential revisit that she will have to serve. On the other hand, the regular doctor knows the patient better and may therefore be able to save the time that a transactional provider would need to spend to elicit the necessary information to obtain the same result.

We explore the duration effect empirically, using the same methods we used for revisit intervals, and report the results in Section 3.11 of Chapter 3. The data suggest that the information benefits of familiarity outweigh the time taken for a more thorough service. Regular doctors spend, on average, less consultation time with their patients than transactional providers. However, the effect size is not so large that a physician could accommodate an additional patient in a typical four-hour clinical session. Nevertheless, these results suggest that increasing care continuity will not necessitate a reduction of daily clinical throughput in primary care practices.

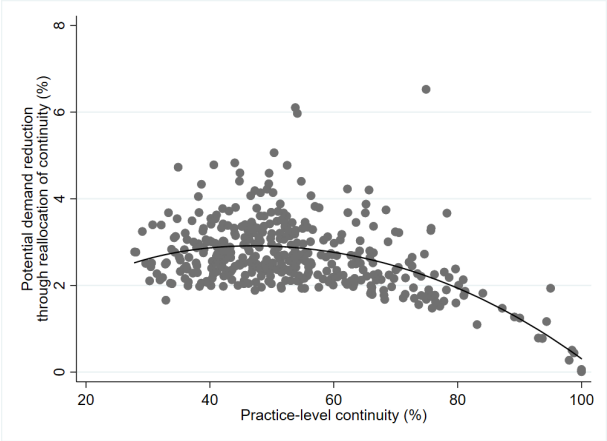

Figure 2.2 Practice-level gains from reallocating consultations with the regular doctor to the most productivity enhancing patients, while keeping the proportion of the practice's consultations with regular doctors unchanged.

2.7 Counterfactual Analysis: Targeting Continuity of Care

Our results suggest that practices could improve productivity by increasing continuity of care and that they could unlock further productivity gains by reallocating continuity to patients who benefit from it the most. In this section, we conduct two analyses to explore retrospectively what the effect would have been on the consultation demand in our data if practices had followed this recommendation in the past.

Using the insights from H2, H2b and H2c, we propose a scoring system that can be used by practice managers to prioritize care continuity, targeting those patients for whom it will have the greatest productivity-enhancing effect. The scoring system ranks consultations by the estimated number of days gained if the consultation is offered by the regular doctor rather than a transactional provider. The estimate is obtained as the difference between the predicted return intervals with and without continuity of care. The prediction is based on Equation Equation 2.1, which we augment by including interactions between the regular doctor variable RD_i and every other covariate in the model. Section 3.14.1 of Chapter 3 provides more details on this estimation approach.

Our first analysis does not assume that practices changed their proportion of continuity consultations. Instead, we only explore the demand reduction they would have achieved had they optimized these consultations by shifting them to the most productivity-enhancing patients, using the above scoring system. Figure 2.2 shows that such targeting of continuity of care has the potential of unlocking productivity gains, even without changing the overall proportion of continuity consultations. If all practices had retained continuity at the same level but better targeted continuity at the most productivity enhancing patients, then the total consultation demand in the sample would have reduced by 2.7%. In fact, as Figure 2.2 shows, some practices would have reduced their demand significantly more.

In our second analysis, we consider what would have happened in our sample if practices not only better targeted care continuity but also increased the level of continuity of care provided. To explore this, we first select a target proportion $0\% \leq x \leq 100\%$ of continuity consultations. We then identify all practices that offered less than $x\%$ continuity of care, and consider the impact on productivity if these practices had offered continuity to $x\%$ of their patients instead. (For those practices offering more than $x\%$ continuity, we leave their proportion of continuity consultations unchanged.) We then re-allocate, as above, the available proportion of continuity consultations in each practice to the most productivity-enhancing patients. This allows us to estimate the counterfactual demand reduction as continuity of care is increased (i.e., by increasing x) and optimally allocated.

The results are summarized in Figure 2.3. If continuity of care levels were increased for those practices with levels below the across-practice median ($x = 51.3\%$ continuity), 75th percentile ($x = 62.3\%$ continuity) and 90th percentile ($x = 74.1\%$ continuity), then total system demand would have reduced by 3.4%, 4.3% and 5.2%, respectively. These estimates show that there are significant productivity gains to be realized by increasing the proportion of patients seeing their regular doctors and by better targeting continuity of care at those patients who benefit from it most.

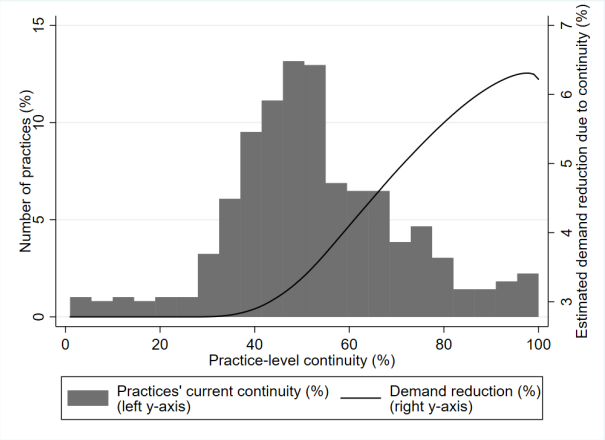

Figure 2.3 System-level reduction in demand if all practices offer a minimum continuity level (specified on the x-axis) to their most productivity-enhancing patients; The underlying histogram shows the current distribution of the consultations with the regular doctor across practices: if all practices were to offer continuity at the across-practice 90th percentile (74.1%), the total system-level demand would have reduced by 5.2%.

2.8 Managerial and Policy Implications

This paper studies the relationship between relational continuity of care and clinical productivity in primary care practices. It complements earlier work on the effect of care continuity on patient outcomes and out-of-practice resource utilization. Based on consultation-level data

from 1,883,626 patients in 381 English primary practices over 11 years, we find evidence that care continuity has a significant positive effect on the productivity of primary care physicians. Specifically, the data shows that revisit intervals for patients are prolonged by an estimated 13.2% (95% CI: $[12.2\%, 14.1\%]$) when patients see their regular doctor instead of a transactional provider, while there is no evidence that consultations with a regular doctor take longer. The positive productivity effect is larger for older patients, patients with long-term medical conditions and patients with mental health problems. We use a comprehensive taxonomy of models to interrogate alternative explanations for causality and find consistent results across all model specifications, suggesting that the correlational effect is unlikely to be caused by confounding or selection bias. Our findings have important operational implications for primary care practice managers as well as strategic implications for the industry and its regulators.

On the operational side, our findings show that there is no intrinsic trade-off between providing continuity of care and maximising the productivity of a practice's clinical workforce. Doubling down on care continuity can be an effective strategy to improve clinical productivity, particularly if a practice operates in a capitation-based funding environment and serves a relatively older and more complex patient population. In fact, since the strength of the continuity-productivity relationship varies significantly across patient groups, practice managers can further improve productivity by targeting specific patients for care continuity, as we have demonstrated in our counterfactual analysis. By contrast, if practices seek to improve productivity by maximising daily throughput per clinician at the cost of continuity of care, they may find that the increase in staffing needed to serve the increased consultation demand outweighs the throughput gains.

Since the analysis focuses on showing that assigning the regular physician to a patient, conditional on the physician being available, improves clinical productivity, it does not consider the trade-off with waiting times if the regular doctor of the patient is not available. To minimize this trade-off, we suggest that practices should continue offering urgent or same-day appointments that are conducted by doctors who are in-charge of urgent appointments on that particular day. Moreover, when there is uncertainty regarding whether the patient should be seen immediately by a transactional provider or wait for the patient to be booked in with his regular doctor, triaging to decide the appropriate action can be delegated to a clinician instead of the receptionist who typically handles such requests. The practices' ability to make the regular doctor available for the appropriate consultations in a timely fashion is a scheduling problem which is beyond the scope of this study.

On the strategic side, our findings are of importance for regulators and third-party payers in relation to the trend in primary care towards faster and more convenient access to general medical expertise. This trend has led to the emergence of at-scale online providers and has accelerated during the COVID-19 pandemic, when patients and doctors were forced to adapt to online consultations. Online primary care providers offer a transactional platform for online appointments, matching on-the-spot demand and supply. Much of the advantage comes from scale and the pooling of clinical time. The service is a single consultation between a doctor and

a patient, not a long-term relationship. Our findings have several implications for this business model, for regulators and payers who wish to create a sustainable primary care environment, and for primary care practices that need to respond strategically to the threat of the new entrants to the market.

First, our findings provide evidence that the online business model of primary care will be particularly profitable in a fee-for-service environment, where patients or third parties pay per consultation, since transactional service provision generates higher future demand than service provision by regular doctors. By contrast, the model is less compelling in a capitation-based or subscription-based funding environment, where providers are paid a fixed fee per patient per month or year and the risk of excess demand is born by the provider.

Second, our analysis of moderators – chronic disease, age, and mental health – shows that in a capitation context, online primary care providers will have a significant incentive to design their services to be less attractive to older, more demanding or more vulnerable patients to minimize the negative productivity effect of transactional medicine for their patients. We believe this will lead to a segmentation of services, with lower-risk patients being increasingly served quickly and conveniently by transaction-focused online providers, while higher-risk patients are served by local practices that offer the relational continuity of care that these patients need and have the close relationship with the local provider network that is required to effectively serve these patients. In fact, most patients will reach a time in their lives when they move from one segment to the other.

If such a service segmentation is the future of primary care, then regulators must anticipate the destabilizing effect that this transformation will have on traditional practices during a transition period. Regulators will need to respond with adequate risk- and productivity-adjusted funding models. The scoring method used in our counterfactual analysis offers a suggestion for how such models could be developed. At the same time, local practice managers face an important strategic choice: Should they provide services for both segments in-house, managing the tensions between the transactional and relational service model, or should they out-source transactional medicine and focus their in-house services on relational medicine for patients who benefit most from continuity of care and from the close integration of their primary care provider in a local provider network that at-scale online providers will find difficult to replicate? More research is required to advise practices on this decision.

References

Abadie A, Spiess J (2021) Robust post-matching inference. *Journal of the American Statistical Association* 1–13.

Ahuja V, Alvarez CA, Staats BR (2020a) How continuity in service impacts variability: Evidence from a primary care setting, SMU Cox School of Business Research Paper No. 19-13.

Ahuja V, Alvarez CA, Staats BR (2020b) Maintaining continuity in service: An empirical examination of primary care physicians. *Manufacturing & Service Operations Management* .

Amjad H, Carmichael D, Austin AM, Chang CH, Bynum JP (2016) Continuity of care and health care utilization in older adults with dementia in fee-for-service medicare. *JAMA internal medicine* 176(9):1371–1378.

Armony M, Roels G, Song H (2021) Pooling queues with strategic servers: The effects of customer ownership. *Operations Research* 69(1):13–29.

Austin PC, Small DS (2014) The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statistics in Medicine* 33(24):4306–4319.

Balint M (1955) The doctor, his patient, and the illness. *The Lancet* 265(6866):683–688.

Barker I, Steventon A, Deeny SR (2017) Association between continuity of care in general practice and hospital admissions for ambulatory care sensitive conditions: cross sectional study of routinely collected, person level data. *Bmj* 356.

Bavafa H, Hitt LM, Terwiesch C (2018) The impact of e-visits on visit frequencies and patient health: Evidence from primary care. *Management Science* 64(12):5461–5480.

Bayliss EA, Ellis JL, Shoup JA, Zeng C, McQuillan DB, Steiner JF (2015) Effect of continuity of care on hospital utilization for seniors with multiple medical conditions in an integrated health care system. *The Annals of Family Medicine* 13(2):123–129.

Beech J, Bottery S, Charlesworth A, Evans H, Gershlick B, Hemmings N, Imison C, Kahtan P, McKenna H, Murray R, Palmer B (2020) Closing the gap: Key areas for action on the health and care workforce — The Nuffield Trust. Technical report, Nuffield Trust.

Benjaafar S (1995) Performance bounds for the effectiveness of pooling in multi-processing systems. *European Journal of Operational Research* 87(2):375–388.

Biringer E, Hartveit M, Sundfør B, Ruud T, Borg M (2017) Continuity of care as experienced by mental health service users-a qualitative study. *BMC Health Services Research* 17(1):763.

Black DA, Smith JA (2004) How robust is the evidence on the effects of college quality? evidence from matching. *Journal of Econometrics* 121(1-2):99–124.

Bobroske K, Freeman M, Huan L, Cattrell A, Scholtes S (2021) Curbing the Opioid Epidemic at its Root: The Effect of Provider Discordance after Opioid Initiation. *Forthcoming in Management Science* .

Brookhart MA, Patrick AR, Schneeweiss S, Avorn J, Dormuth C, Shrank W, van Wijk BL, Cadarette SM, Canning CF, Solomon DH (2007) Physician follow-up and provider continuity are associated with long-term medication adherence: a study of the dynamics of statin use. *Archives of internal medicine* 167(8):847–852.

Cai X, Gong J, Lu Y, Zhong S (2018) Recover overnight? work interruption and worker productivity. *Management Science* 64(8):3489–3500.

Calmon AP, Ciocan FD, Romero G (2021) Revenue management with repeated customer interactions. *Management Science* 67(5):2944–2963.

Chen HM, Tu YH, Chen CM (2017) Effect of continuity of care on quality of life in older adults with chronic diseases: a meta-analysis. *Clinical Nursing Research* 26(3):266–284.

Cho KH, Kim YS, Nam CM, Kim TH, Kim SJ, Han KT, Park EC (2015) The association between continuity of care and all-cause mortality in patients with newly diagnosed obstructive pulmonary disease: a population-based retrospective cohort study, 2005-2012. *PloS one* 10(11).

Chun SY, Iancu DA, Trichakis N (2020) Loyalty program liabilities and point values. *Manufacturing & Service Operations Management* 22(2):257–272.

Chun SY, Ovchinnikov A (2019) Strategic consumers, revenue management, and the design of loyalty programs. *Management Science* 65(9):3969–3987.

Dall T, Reynolds R, Chakrabarti R, Jones K, Iacobucci W (2020) The Complexities of Physician Supply and Demand: Projections from 2018 to 2033. Technical Report June, Association of American Medical Colleges.

Diamond DW (1991) Monitoring and reputation: The choice between bank loans and directly placed debt. *Journal of political Economy* 99(4):689–721.

Dossa AR, Moisan J, Guénette L, Lauzier S, Grégoire JP (2017) Association between interpersonal continuity of care and medication adherence in type 2 diabetes: an observational cohort study. *CMAJ open* 5(2):E359.

Drury A, Payne S, Brady AM (2020) Identifying associations between quality of life outcomes and healthcare-related variables among colorectal cancer survivors: A cross-sectional survey study. *International Journal of Nursing Studies* 101:103434.

Evans P, Langley P, Gray DP (2008) Diagnosing type 2 diabetes before patients complain of diabetic symptoms—clinical opportunistic screening in a single general practice. *Family practice* 25(5):376–381.

Freeman G, Hughes J, et al. (2010) Continuity of care and the patient experience. Technical report, The King's Fund.

Freeman M, Robinson S, Scholtes S (2020) Gatekeeping, fast and slow: An empirical study of referral errors in the emergency department. *Management Science* .

Goodwin N, Curry N, Naylor C, Ross S, Duldig W, et al. (2010) Managing people with long-term conditions. *London: The Kings Fund* .

Gubler T, Larkin I, Pierce L (2018) Doing well by making well: The impact of corporate wellness programs on employee productivity. *Management Science* 64(11):4967–4987.

Haggerty JL, Reid RJ, Freeman GK, Starfield BH, Adair CE, McKendry R (2003) Continuity of care: a multidisciplinary review. *Bmj* 327(7425):1219–1221.

Hallvik SE, Geissert P, Wakeland W, Hildebran C, Carson J, O'kane N, Deyo RA (2018) Opioid-prescribing continuity and risky opioid prescriptions. *The Annals of Family Medicine* 16(5):440–442.

Henderson BJ, Tookes H (2012) Do investment banks' relationships with investors impact pricing? the case of convertible bond issues. *Management Science* 58(12):2272–2291.

Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, Van Staa T, Smeeth L (2015) Data Resource Profile Data Resource Profile: Clinical Practice Research Datalink (CPRD). *International Journal of Epidemiology* 827–836.

Hill AP, Freeman GK (2011) Promoting continuity of care in general practice. *London: Royal College of General Practitioners* .

Hjortdahl P, Borchgrevink CF (1991) Continuity of care: influence of general practitioners' knowledge about their patients on use of resources in consultations. *British Medical Journal* 303(6811):1181–1184.

Huntley A, Lasserson D, Wye L, Morris R, Checkland K, England H, Salisbury C, Purdy S (2014) Which features of primary care affect unscheduled secondary care use? a systematic review. *BMJ open* 4(5):e004746.

Institute for Government (2019) General practice — The Institute for Government.

Jann B (2006) GSAMPLE: Stata module to draw a random sample. Statistical Software Components, Boston College Department of Economics.

Jeffers H, Baker M (2016) Continuity of care: still important in modern-day general practice.

Jouini O, Dallery Y, Nait-Abdallah R (2008) Analysis of the impact of team-based organizations in call center management. *Management Science* 54(2):400–414.

Kajaria-Montag H, Freeman M (2020) Explaining the erosion of relational care continuity: An empirical analysis of primary care in England. *SSRN Electronic Journal* INSEAD Working Paper No. 2020/47/TOM.

Katz DA, McCoy KD, Vaughan-Sarrazin MS (2015) Does greater continuity of veterans administration primary care reduce emergency department visits and hospitalization in older veterans? *Journal of the American Geriatrics Society* 63(12):2510–2518.

Knaak S, Mantler E, Szeto A (2017) Mental illness-related stigma in healthcare: Barriers to access and care and evidence-based solutions. *Healthcare management forum*, volume 30, 111–116 (SAGE Publications Sage CA: Los Angeles, CA).

Koopman RJ, Mainous III AG, Baker R, Gill JM, Gilbert GE (2003) Continuity of care and recognition of diabetes, hypertension, and hypercholesterolemia. *Archives of Internal Medicine* 163(11):1357–1361.

Leniz J, Gulliford MC (2019) Continuity of care and delivery of diabetes and hypertensive care among regular users of primary care services in chile: a cross-sectional study. *BMJ Open* 9(10).

Li MM, Nassiri S, Liu X, Ellimoottil C (2021) How does telemedicine shape physician's practice in mental health? *Shima and Liu, Xiang and Ellimoottil, Chandy, How Does Telemedicine Shape Physician's Practice in Mental Health* .

Lin IP, Wu SC, Huang ST (2015) Continuity of care and avoidable hospitalizations for chronic obstructive pulmonary disease (copd). *The Journal of the American Board of Family Medicine* 28(2):222–230.

Liu N, Finkelstein SR, Kruk ME, Rosenthal D (2018) When waiting to see a doctor is less irritating: Understanding patient preferences and choice behavior in appointment scheduling. *Management Science* 64(5):1975–1996.

Maarsingh OR, Henry Y, van de Ven PM, Deeg DJ (2016) Continuity of care in primary care and association with survival in older people: a 17-year prospective cohort study. *British Journal of General Practice* 66(649):e531–e539.

Millimet DL, Tchernis R (2013) Estimation of treatment effects without an exclusion restriction: With an application to the analysis of the school breakfast program. *Journal of Applied Econometrics* 28(6):982–1017.

NHS Digital (2015) NHS Payments to General Practice, England 2014/15. Technical Report September, NHS Digital.

Nyweide DJ, Anthony DL, Bynum JP, Strawderman RL, Weeks WB, Casalino LP, Fisher ES (2013) Continuity of care and the risk of preventable hospitalization in older adults. *JAMA Internal Medicine* 173(20):1879–1885.

O'Connor PJ, Desai J, Rush WA, Cherney LM, Solberg LI, Bishop DB (1998) Is having a regular provider of diabetes care related to intensity of care and glycemic control? *Journal of Family Practice* 47:290–297.

Özer Ö, Subramanian U, Wang Y (2018) Information sharing, advice provision, or delegation: what leads to higher trust and trustworthiness? *Management Science* 64(1):474–493.

Özer Ö, Zheng Y, Ren Y (2014) Trust, trustworthiness, and information sharing in supply chains bridging china and the united states. *Management Science* 60(10):2435–2460.

Palmer W (2019) Is the number of GPs falling across the UK? Technical report, Nuffield Trust, UK.

Palmer W, Hemmings N, Rosen R, Keeble E, Williams S, Imison C (2018) Improving access and continuity in general practice. *Research Summary* .

Peel MJ (2018) Addressing unobserved selection bias in accounting studies: The bias minimization method. *European Accounting Review* 27(1):173–183.

Pierce L, Snow DC, McAfee A (2015) Cleaning house: The impact of information technology monitoring on employee theft and productivity. *Management Science* 61(10):2299–2319.

Popescu I, Wu Y (2007) Dynamic pricing strategies with reference effects. *Operations research* 55(3):413–429.

Pourat N, Davis AC, Chen X, Vrungos S, Kominski GF (2015) In california, primary care continuity was associated with reduced emergency department use and fewer hospitalizations. *Health Affairs* 34(7):1113–1120.

Queenan C, Cameron K, Snell A, Smalley J, Joglekar N (2019) Patient heal thyself: reducing hospital readmissions with technology-enabled continuity of care and patient activation. *Production and Operations Management* 28(11):2841–2853.

Ramanayake R, Basnayake B (2018) Evaluation of red flags minimizes missing serious diseases in primary care. *Journal of Family Medicine and Primary Care* 7(2):315.

Ranganathan A, Shivaram R (2021) Getting their hands dirty: How female managers motivate female worker productivity through subordinate scut work. *Management Science* 67(5):3299–3320.

Reagans R, Argote L, Brooks D (2005) Individual experience and experience working together: Predicting learning rates from knowing who knows what and knowing how to work together. *Management science* 51(6):869–881.

Reuter J (2006) Are IPO allocations for sale? Evidence from mutual funds. *The Journal of Finance* 61(5):2289–2324.

Ride J, Kasteridis P, Gutacker N, Doran T, Rice N, Gravelle H, Kendrick T, Mason A, Goddard M, Siddiqi N, et al. (2019) Impact of family practice continuity of care on unplanned hospital use for people with serious mental illness. *Health Services Research* 54(6):1316–1325.

Rosen R, Massey Y, Abbas S, Hufflett T (2020) Relational continuity for general practice patients with new and changing symptoms. Technical report, Valentine Health Partnership, The Health Foundation.

Rosenbaum PR (2005) Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *The American Statistician* 59(2):147–152.

Rothkopf MH, Rech P (1987) Perspectives on queues: Combining queues is not always beneficial. *Operations Research* 35(6):906–909.

Royal College of General Practitioners (2019) Fit for the future: a vision for general practice.

Rubinovitch M (1985) The slow server problem. *Journal of Applied Probability* 205–213.

Sampson F, Pickin M, O'Cathain A, Goodall S, Salisbury C (2008) Impact of same-day appointments on patient satisfaction with general practice appointment systems. *British Journal of General Practice* 58(554):641–643.

Saparito PA, Chen CC, Sapienza HJ (2004) The role of relational trust in bank–small firm relationships. *Academy of Management Journal* 47(3):400–410.

Semahegn A, Torpey K, Manu A, Assefa N, Tesfaye G, Ankomah A (2018) Psychotropic medication non-adherence and associated factors among adult patients with major psychiatric disorders: a protocol for a systematic review. *Systematic Reviews* 7(1):1–5.

Senot C (2019) Continuity of care and risk of readmission: An investigation into the healthcare journey of heart failure patients. *Production and Operations Management* 28(8):2008–2030.

Smith DR, Whitt W (1981) Resource sharing for efficiency in traffic systems. *Bell System Technical Journal* 60(1):39–55.

Song H, Tucker AL, Murrell KL (2015) The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science* 61(12):3032–3053.

Song H, Tucker AL, Murrell KL, Vinson DR (2018) Closing the productivity gap: Improving worker productivity through public relative performance feedback and validation of best practices. *Management Science* 64(6):2628–2649.

Sosa ME (2011) Where do creative interactions come from? the role of tie content and social networks. *Organization Science* 22(1):1–21.

Staats BR, Gino F (2012) Specialization and variety in repetitive tasks: Evidence from a japanese bank. *Management science* 58(6):1141–1159.

Sunar N, Tu Y, Ziya S (2021) Pooled vs. dedicated queues when customers are delay-sensitive. *Management Science* .

Tammes P, Purdy S, Salisbury C, MacKichan F, Lasserson D, Morris RW (2017) Continuity of primary care and emergency hospital admissions among older patients in england. *The Annals of Family Medicine* 15(6):515–522.

Tan TF, Netessine S (2014) When does the devil make work? an empirical study of the impact of workload on worker productivity. *Management Science* 60(6):1574–1593.

Tarrant C, Dixon-Woods M, Colman AM, Stokes T (2010) Continuity and trust in primary care: a qualitative study informed by game theory. *The Annals of Family Medicine* 8(5):440–446.

Taylor TA, Plambeck EL (2007) Supply chain relationships and contracts: The impact of repeated interaction on capacity investment and procurement. *Management science* 53(10):1577–1593.

The King's Fund (2016) Understanding pressures in general practice — The Kings Fund. Technical report, The King's Fund.

von Bültzingslöwen I, Eliasson G, Sarvimäki A, Mattsson B, Hjortdahl P (2006) Patients' views on interpersonal continuity in primary care: a sense of security based on four core foundations. *Family practice* 23(2):210–219.

Wang J, Zhou YP (2018) Impact of queue configuration on service time: Evidence from a supermarket. *Management Science* 64(7):3055–3075.

Wilkin D, Metcalfe D (1984) List size and patient contact in general medical practice. *British Medical Journal (Clinical Research Ed.)* 289(6457):1501–1505.

Wooldridge JM (2002) Econometric analysis of cross section and panel data mit press. *Cambridge, MA* 108.

Wooldridge JM (2015) Control function methods in applied econometrics. *Journal of Human Resources* 50(2):420–445.

Worrall G, Knight J (2011) Continuity of care is good for elderly people with diabetes: retrospective cohort study of mortality and hospitalization. *Canadian Family Physician* 57(1):e16–e20.

Ye T, Sun X, Tang W, Miao Y, Zhang Y, Zhang L (2016) Effect of continuity of care on health-related quality of life in adult patients with hypertension: a cohort study in china. *BMC Health Services Research* 16(1):674.

Chapter 3

Continuity of care increases clinical productivity – Further investigations

3.1 Introduction

This chapter includes supporting material designed to go with the analysis presented in Chapter 2. Therefore, it is not meant to be read in isolation but rather serves as material that provides more background and support for the methods, techniques and results presented, as well as helps to confirm the robustness of the main findings.

3.2 Sample inclusion criteria

Below we provide justification for the various inclusion criteria that were used to transform the initial data set into the final sample for analysis. Each of these points is also summarized in Table 2.1 in Section 2.4.2 of Chapter 2.

- At most practices, patients are able to see a range of staff, such as physicians, nurses, and nurse practitioners, through various channels including face-to-face and phone consultations. In accordance with the literature on continuity of care, we include only consultations with physicians; there is currently only weak evidence of advantages of continuity between patients and non-physician care providers (Tammes et al. 2017, Barker et al. 2017). We also include only face-to-face consultations, which were the most common way patients interacted with their physicians.

- To ensure high data quality, CPRD determines when the data from a practice is considered to be of sufficient quality for research purposes. We discard any observations made before the date at which a practice's data is considered to be of research quality.

- Patients change registrations between practices over time, for example when they change their place of residency. To ensure that registration gaps with a practice do not affect the analysis, we only include the latest continuous period of registration of a patient with a practice.

- The data allow us to distinguish between two types of physicians, *established* physicians who have a contract with the practice and *unestablished* physicians who are not permanent employees of the practice. Instead they may be self-employed or employed with an agency. They work on an ad-hoc basis, often in multiple practices, and are paid on an hourly basis.[1,2]

- To increase the homogeneity of the sample, we have excluded babies and children as they have different medical needs to adults. The sample thus includes only consultations associated with adult patients over the age of 18.

- In line with previous literature, we exclude consultations with patients who had fewer than three consultations in the two-year window prior to the consultation since an accurate identification of a regular doctor is not possible with very few consultations (Ahuja et al. 2020). We also exclude consultations that were preceded by more than 104 consultations over a two-year window, an average of one consultation per week. Such patients are likely to have very special needs or are on a specific care management plan that requires frequent planned visits. The relationship between care continuity and the revisit interval is less meaningful for this group of patients.

- Since we use a two-year period to calculate the patient's regular doctor, we exclude consultations that took place in the first two years following the patient's registration date at the practice. During this period, we do not have an accurate estimate of the patient's regular doctor.

- We exclude a patient's last recorded consultation as there is no subsequent consultation with which to calculate the revisit interval. We refer to the remaining consultations as consultations with a valid revisit interval.

- Our data consists of the complete medical record of all patients described above who had a contact with a primary care practice between January 2007 and December 2017[3]. However, we also have partial records on some patients outside of this date range, which we remove from the analysis.

- We wish to estimate the effect of care continuity at times when the practice is able to schedule an appointment with the patient's regular doctor. We therefore exclude consultations that take place during weeks when the focal patient's regular doctor is on leave. Our results should therefore be interpreted as being conditional on the patient's regular doctor being available in the week during which the consultation occurred.

[1]From the 67 possible staff roles coded in the CPRD data, we classify GPs as defined by Vision (GPs, the UK equivalent of a PCP in the US)– the EMR system used by the practices in our sample.

[2]Type of consultations are coded into 51 different types of consultations, for which we use the classification system from Kontopantelis et al. (2015) to categorize face-to-face visits.

[3]This restriction is applied by CPRD, our data provider

- Finally, due to statistical estimation issues related to the large sample size, all of the models are estimated on a random sample of 25% of the remaining 45,376,070 consultations. The random samples are drawn from each of the 381 primary care practices remaining in the data and then merged, thus ensuring that each practice continues to be represented.

3.3 What predicts continuity of care? Validation of the independent variable

To validate that we are capturing continuity correctly in our independent variable, we wish to understand the relationship between patient-level factors and the probability that the patient will see their regular doctor. We expect that the patient's past frequency of consultations and the patient's complexity (age, number of comorbidities, etc.) will be the most important predictors. To investigate this, we estimate a linear probability model of the form $RD_i = \beta_0 + \boldsymbol{\beta}_1 \mathbf{X}$, where X is a vector of control variables described in Section 2.4.3.4 and RD_i is a binary variable, capturing whether the patient sees her regular doctor. The results of the model are given in Table 3.1.

	Dependent Variable: RD_i	
$ln(ExpectedRI)$	-0.026***	[-0.026,-0.025]
IMD=2	-0.007***	[-0.009,-0.006]
IMD=3	-0.008***	[-0.009,-0.006]
IMD=4	-0.017***	[-0.018,-0.015]
IMD=5	-0.020***	[-0.022,-0.019]
Demand	-0.025***	[-0.027,-0.024]
Female	-0.025***	[-0.026,-0.024]
26-30 yrs	0.017***	[0.015,0.019]
31-35 yrs	0.034***	[0.032,0.036]
36-40 yrs	0.050***	[0.048,0.052]
41-15 yrs	0.070***	[0.069,0.072]
46-50 yrs	0.086***	[0.084,0.088]
51-55 yrs	0.101***	[0.099,0.103]
56-60 yrs	0.115***	[0.113,0.117]
61-65 yrs	0.127***	[0.125,0.129]
66-70 yrs	0.140***	[0.138,0.142]
71-75 yrs	0.150***	[0.148,0.152]
76-80 yrs	0.155***	[0.152,0.157]
81-85 yrs	0.150***	[0.147,0.152]
86+ yrs	0.132***	[0.130,0.135]
1 comorbidity	0.028***	[0.026,0.029]
2 comorbidites	0.037***	[0.035,0.039]
3 comorbidites	0.040***	[0.037,0.043]
4 comorbidites	0.044***	[0.040,0.048]
$\geq$ 5 comorbidites	0.049***	[0.044,0.055]
1 prescription	0.022***	[0.021,0.023]
2 prescriptions	0.033***	[0.032,0.035]
3 prescriptions	0.043***	[0.041,0.044]
4-5 prescriptions	0.049***	[0.048,0.051]
6-7 prescriptions	0.052***	[0.051,0.054]
8-9 prescriptions	0.053***	[0.051,0.055]
10-12 prescriptions	0.052***	[0.050,0.054]
13-15 prescriptions	0.053***	[0.050,0.055]
16+ prescriptions	0.059***	[0.055,0.062]
4 consultations	0.008***	[0.007,0.010]
5 consultations	0.012***	[0.010,0.013]
6 consultations	0.007***	[0.005,0.008]
7 consultations	0.009***	[0.008,0.011]
8 consultations	0.011***	[0.009,0.012]
9 consultations	0.014***	[0.012,0.015]
10 consultations	0.015***	[0.013,0.017]
11-12 consultations	0.016***	[0.015,0.018]
13-14 consultations	0.020***	[0.018,0.022]
15-16 consultations	0.021***	[0.019,0.022]
17-18 consultations	0.022***	[0.020,0.024]
19-20 consultations	0.023***	[0.021,0.025]
21-23 consultations	0.024***	[0.022,0.026]
24-26 consultations	0.026***	[0.024,0.028]
27-30 consultations	0.026***	[0.024,0.028]
31-35 consultations	0.026***	[0.023,0.028]
36-45 consultations	0.027***	[0.024,0.029]
46-55 consultations	0.034***	[0.030,0.038]
56+ consultations	0.046***	[0.042,0.051]
Established GP=1	0.023***	[0.021,0.024]
Mental Health=1	0.020***	[0.019,0.021]
Constant	0.616***	[0.607,0.624]
Observations	11,344,065	
All other controls	Yes	

95% confidence intervals in brackets

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3.1 Linear probability model to validate the independent variable

We find that the older the patient is, the higher the probability that they will see their regular doctor. Similarly, the more complex the patient is (higher comorbidities and higher number of prescriptions), the higher the probability that they will see their regular doctor. We also find an increasing relationship with the patient's past frequency of consultation – the probability of seeing the regular doctor increases with the increase in past frequency. However, further analysis shows us that this increase is active only up to a certain point (60 consultations), after which the probability of seeing the regular doctor stabilizes (Figure 3.1).

As the observed effects in Table 3.1 are in line with expectations, we have confidence that the binary variable RD_i is appropriately identifying the doctor who is more likely to be a patients regular doctor.

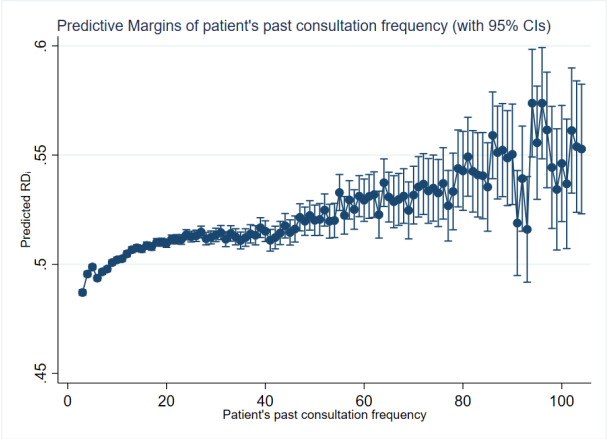

Figure 3.1 Non-linear effect of the patient's past consultation frequency on whether the patient saw her regular doctor

3.4 Alternative measures of continuity of care

We also use different measures of continuity of care as specified in the literature to investigate the effect on the revisit intervals and to provide robustness for our results. First, we use the Bice-Boxerman Continuity of Care index (COCI) which represents the dispersion of visits between providers or the degree of coordination needed between different providers (Pollack et al. 2016). This measure is suitable for longitudinal data and provides comparability across patients. The COCI as calculated as follows:

$$COCI = \frac{(\sum_{i=1}^{p} n_i^2) - n}{n(n-1)}$$

where n_i is the number of visits to provider i and n is the total number of visits. For our setting, we calculate the COCI for every consultation by using all consultations that took place in the two-year interval preceding the focal consultation.

The second measure we use is the Herfindahl–Hirschman index (HHI) which is conceptually similar to the COCI but measures the extent to which the patient's visits are concentrated with a single or group of providers (Maarsingh et al. 2016, Pollack et al. 2016). The HHI is most commonly used to measure market concentration in economic analyses but is also now widely used in the medical literature. The HHI is calculated as follows:

$$HHI = \sum_{i=1}^{p} (\frac{n_i}{n})^2$$

where n_i is the number of visits to provider i and n is the total number of visits. For our setting, we calculate the HHI for every consultation by using all consultations that took place in the two-year interval preceding the focal consultation.

The third measure we use is the Usual Provider of Care index (UPC) which measures the density or the concentration of visits to a single usual provider. For each consultation, the UPC is the number of visits to the regular provider divided by the total number of visits (Maarsingh et al. 2016).

Table 3.2 below shows the effect of COCI, HHI and UPC on the revisit intervals. The results suggest that a higher COCI (lower dispersion), higher HHI (lower dispersion) and a higher UPC (higher density with one provider), all lead to a longer revisit interval.

Table 3.2 Coefficient of RD_i, UPC, HHI and COCI.

Dependent Variable = Natural logarithm of the revisit interval				
$RD_i = 1$	0.140^{***}			
	[0.138,0.141]			
UPC		0.294^{***}		
		[0.290,0.298]		
HHI			0.149^{***}	
			[0.145,0.153]	
COCI				0.122^{***}
				[0.118,0.126]
Constant	3.594^{***}	3.703^{***}	3.582^{***}	3.624^{***}
	[3.571,3.617]	[3.680,3.727]	[3.560,3.605]	[3.601,3.647]
Observations	11,344,065	11,344,065	11,344,065	11,344,065
Adjusted R^2	0.144	0.144	0.142	0.142

95% confidence intervals in brackets

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: 95% confidence intervals in square brackets

3.5 Measurement of continuity related concerns and preliminary evidence

In the way our variables are constructed, there may be certain measurement issues if the patient's regular doctor is away (for e.g, if the doctor is on holiday, extended leave, moved, retired or left the practice). A patient who visits primary care less frequently is more likely to be assigned to a regular doctor who is unavailable during a certain week. Hence, the less frequently a patient visits, the more consultations it takes our doctor assignment algorithm to identify an "updated" regular doctor for the patient after their original regular doctor is away.

To address this issue, we identify every instance when the doctor is away for at least one full week, select the first week from this period for each doctor and refer to this as week w. We then compare the average RI of patients during week w as compared to $w-1$. During week w none of the patients are able to see their regular doctor and care continuity is 0 for all patients. During week $w-1$ some patients see their regular doctor and some don't. As outlined in Table 3.3, we find that the average RI for patients who see their regular doctor in week $w-1$ is 3.47. Whereas, for patients who do not see their regular doctor, the average RI across week w and week $w-1$ is not significantly different (3.34 in week $w-1$ compared to 3.32 in week w). Comparing a RI of 3.32 with 3.47 yields a log-increase of 0.15, which provides suggestive preliminary evidence that seeing the regular doctor increases the RI by about 15%.

We repeat this analysis for the consultation duration, and find that the average consultation duration for patients who see their regular doctor in week $w-1$ is 9.09 minutes. For patients who do not see their regular doctor, the average consultation duration across week w and week $w-1$ is 9.29 minutes and 9.37 minutes respectively, and is higher than when the patient sees his regular doctor. Comparing the consultation duration across these scenarios, we find that the effect of seeing the regular doctor decreases the consultation duration by approximately between 3%, in other words transactional providers spend slightly longer on average with patients.

Due to a relatively low impact of care continuity on duration, we focus our main analysis on the revisit intervals.

Table 3.3 Average RI and consultation duration of patients when the regular doctor is present (week $w-1$) compared to when the doctor is away (week w)

	doctor present($w-1$)	doctor present($w-1$)	doctor away(w)
RD_i	0	1	0
N	339,721	194,742	423,992
(1) Revisit Interval (RI)	3.34	3.47	3.32
(2) Duration (mins)	9.37	9.09	9.29

3.6 Description of control variables

To account for various factors that may confound the relationship between continuity of care and revisit intervals, we include different control variables.

Patient demographics

First, differences in patient demographics may affect the patient's preference for seeing her regular doctor as well as her primary care consultation frequency and hence the revisit interval. To account for this, we control for patient age, as older patients are likely to be sicker and make more frequent contact with primary care, and they are also more likely to prefer to see their regular doctor. We include patient age at the time of consultation in categories as shown in Figure 3.2 (Left).

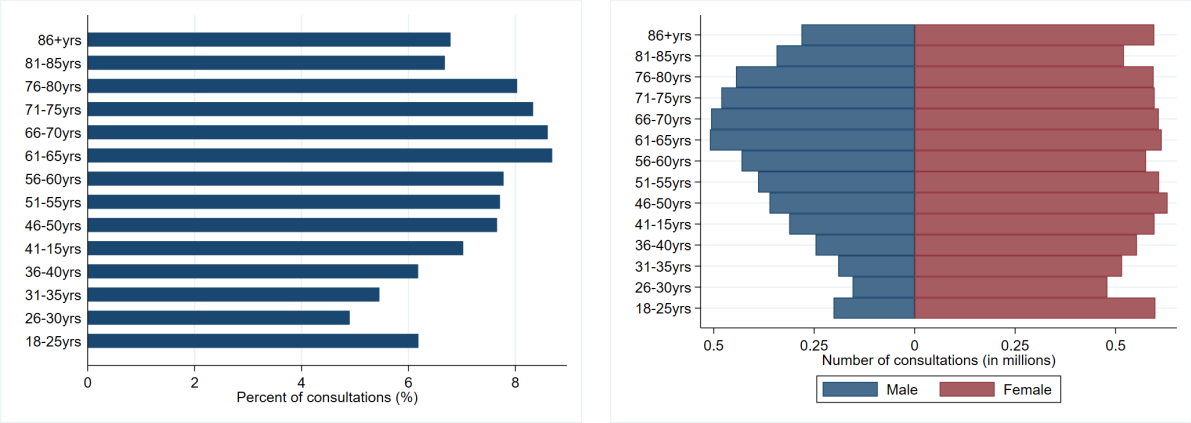

Figure 3.2 (Left) Categories of patient age at the time of consultation; (Right) Breakdown of consultations by gender and age.

Similarly, using the Cambridge Comorbidity Index (CCI), we control for (i) the total number of chronic conditions the patient suffers from at the time of consultation, (ii) a binary variable for each of the 26 individual comorbidities in the CCI and (iii) whether the patient suffers from a mental health condition (described as anxiety, depression or schizophrenia)[4]. The categories used for the total number of chronic conditions the patient suffers from at the time of consultation and the prevalence of the comorbidities in our dataset are represented visually in Figure 3.3.

We control for gender as it might be a potential confounder – for example, pregnant women or women who just gave birth are likely to visit a primary care provider at short intervals and also more likely to want to see a doctor they trust. The breakdown of gender by age group is shown in Figure 3.2 (Right) and suggests that for all age groups, women are heavier users of primary care than men.

[4]There are a total of 38 comorbidities specified in the Cambridge Comorbidity Index, out of which we group 3 comorbidities as mental health condition, and only include those remaining conditions that have a higher than 1% prevalence in our dataset. The leaves us with 26 conditions that we include in the analysis (Payne et al. 2020).

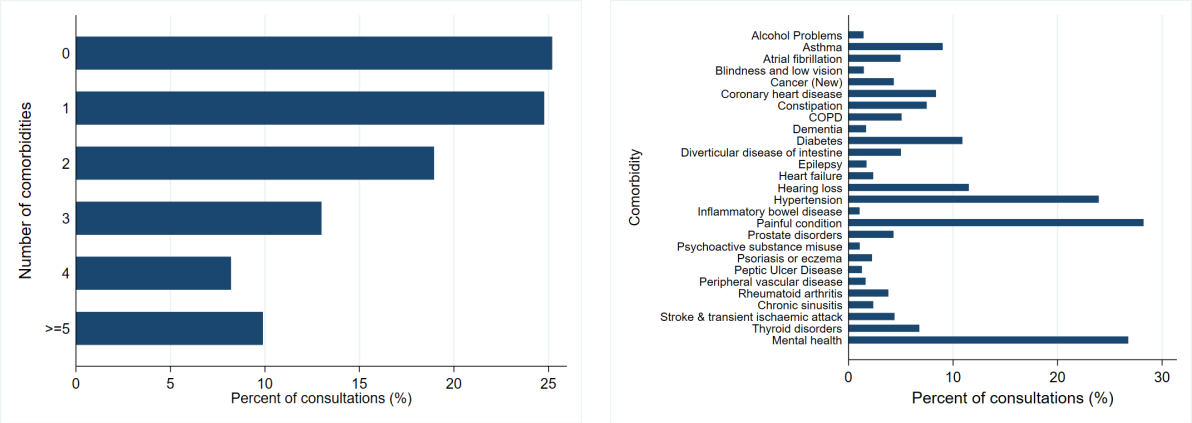

Figure 3.3 (Left) Categories of the number of active comorbidities the patient suffers from at the time of consultation; (Right) Prevalence of each individual comorbidity in the data.

To further capture the patient's severity and complexity, we calculate the total number of unique repeat medications the patient has been prescribed in the 6-month window preceding the focal consultation. The categories used for the total number of prescriptions is represented visually in Figure 3.4.

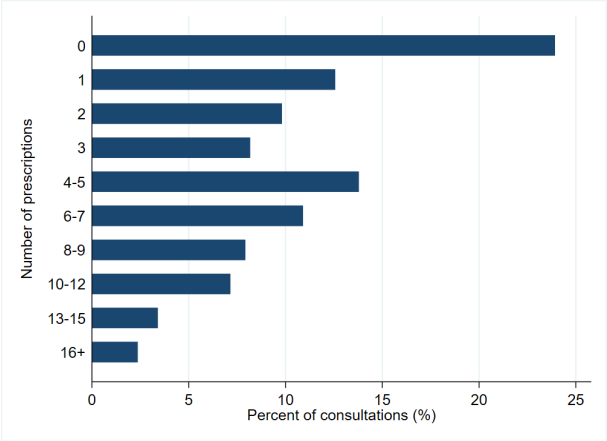

Figure 3.4 Categories of the number of different active repeat prescriptions the patient is prescribed within a 6-month window preceding the consultation

Finally, we control for the socioeconomic status of the patient using the patient-level index of multiple deprivation (IMD), which is provided by CPRD in quintiles and is a relative characterization of poverty or the socioeconomic situation of the area where the patient resides. The categories of the IMD are shown in Figure 3.5.

Patient visit history

An important confounding factor that need to be accounted for is the patient's visit history. The more frequently a patient visited in the past, the more likely they are to also visit frequently in the future. Hence, we expect that such a patient will have a shorter revisit interval. At

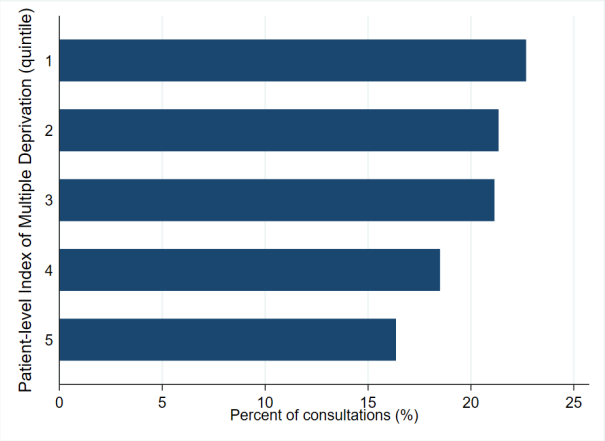

Figure 3.5 Patient's Index of Multiple Deprivation (IMD) as quintiles

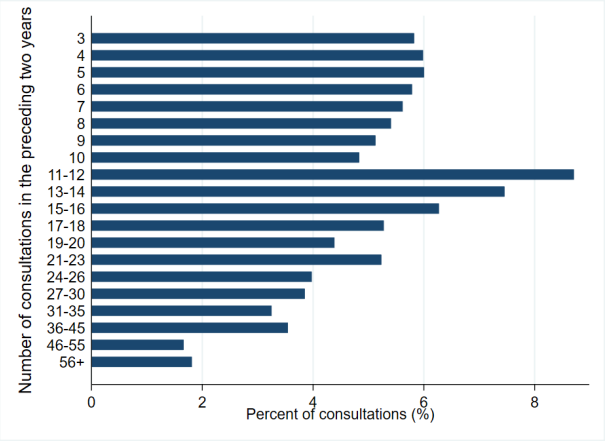

Figure 3.6 Categories of patient's consultations in the two years preceding the focal consultation

the same time, a patient's likelihood of a consultation with a regular doctor is affected by her consultation frequency (see Figure 3.1). Since we anticipate nonlinear effects (see Figure 3.1), we include both the patient's past consultation frequency as a categorical control as well as her average past revisit interval as a linear continuous control (see Table 2.3 of Chapter 2 for details). The categories are shown in Figure 3.6.

Doctor type

Third, we note that practices are staffed with two types of GPs:

- **Established:** Partners, who are co-owners of the practice, and salaried doctors, who are permanent employees of the practice. Usually, established doctors are list-holding doctors who are accountable for the health of their patient list.
- **Unestablished:** Locums or temporary doctors who have temporary contracts with the practice and work ad-hoc shifts. Unestablished doctors are either self-employed or employed at a locum agency, and they are paid on an hourly basis.

Even though the two categories of doctors receive similar training, the incentives for the established doctors to "get it right the first time" might be stronger, which would translate to a higher quality and productivity benefit if the patient's regular provider is an established doctor. According to our regular provider assignment algorithm, established doctors are regular providers of the patient 93% of the time, whereas unestablished doctors are regular providers the rest of the time.

Practice and time-related controls

Fourth, we include practice fixed effects to capture unobserved time-invariant heterogeneity across practices. For practice-level observed time variant heterogeneity, we include a dynamic control that measures practice-level variation in demand. We calculate this as the total practice demand during the focal week of each consultation as compared to a weekly average in a 52-week period around the focal week. This measure captures fluctuations in demand, which may be correlated with the probability that the patient will see her regular doctor and with the scheduling capacity of the practice.

Lastly, time fixed effects are included to account for any factors that change over time and have a common effect on all practices. Controls are included for year to account for trends, month of year to account for seasonality, and day of the week effects to account for differences in demand and supply across different days of the week that may affect both a patient's seeing her regular doctor and the revisit interval.

3.7 Patient visits

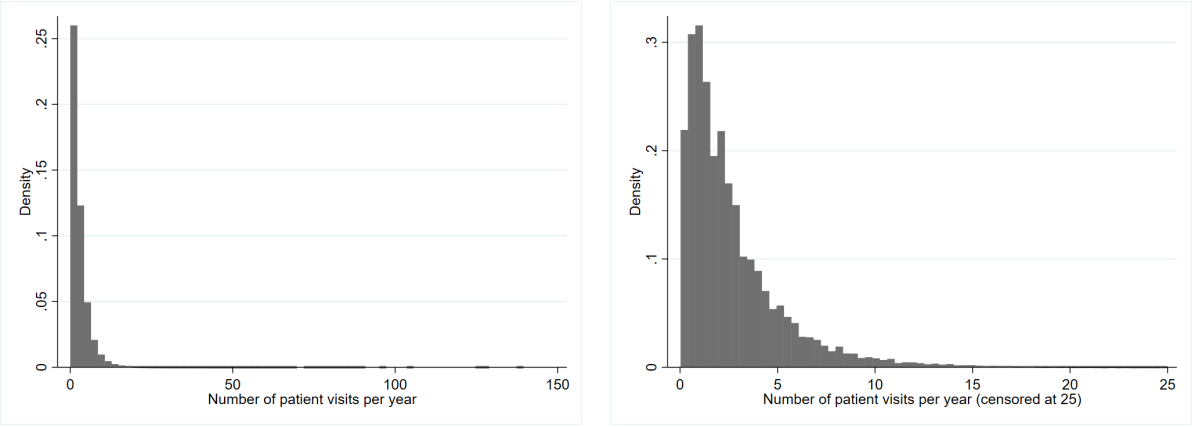

Figure 3.7 (Left) Histogram of number of patient visits per year; (Right) Histogram of number of patient visits per year (censored at 25 visits)

In Figure 3.7 we show the histogram of patient visits in our data. For each patient, we calculate the total number of visits in the data divided by the total number of years registered in the

data. We find that the average number of visits per year is 2.7, the median number of visits is 2.0, with a standard deviation of 2.8 visits.

3.8 Instrumental variable construction

To calculate the instrumental variable, let C_{gt} denote the set of consultations in week t for which doctor g is specified as patients' regular doctor. In other words, $|C_{gt}|$ denotes the total number of consultations in week t made by patients for whom doctor g is their regular doctor, where $|\cdot|$ specifies the cardinality of the set. Let g_c denote the actual doctor who the patient saw during consultation c. Then we define the accessibility of doctor g during week t for patients who consider doctor g as their regular doctor to be

$$WkAccessibility_{gt} = \frac{\sum_{c \in C_{gt}} I\left[g_c = g\right]}{|C_{gt}|}$$

where $I\left[\cdot\right]$ is an indicator function that takes value 1 when the condition inside the brackets is satisfied, and 0 otherwise.

Note that $WkAccessibility_{gt}$ also includes any visits by patient i, which may produce a mechanical relationship between the instrumental variable and the dependent variable in the selection equation (i.e., whether or not the patient saw their regular doctor). To prevent this, let C_{igt} be the set of consultations made by patient i in week t for which doctor g was specified as their regular doctor. Then we can define accessibility, excluding patient i, as follows:

$$WkAccessibility_{igt} = \frac{\sum_{c \in \{C_{gt} \backslash C_{igt}\}} I\left[g_c = g\right]}{|\{C_{gt} \backslash C_{igt}\}|}$$

Next, we standardize $WkAccessibility_{igt}$ by calculating the same measure over a 52-week period preceding week t. Specifically, let $T_t = \{t-52, \ldots, t-2, t-1\}$ denote the set of 52 weeks prior to week t, excluding week t itself. Then the average accessibility over this 52-week period can be written as

$$YrAccessibility_{igt} = \frac{\sum_{t \in T_t} \sum_{c \in \{C_{gt} \backslash C_{igt}\}} I\left[g_c = g\right]}{\sum_{t \in T_t} |\{C_{gt} \backslash C_{igt}\}|}$$

Finally, we define the instrumental variable, IV, as:

$$IV_{igt} = \frac{WkAccessibility_{igt}}{YrAccessibility_{igt}}$$

3.8.1 Statistical tests of the instrumental variable

To validate the instrumental variable, we perform formal hypothesis test for under- and weak identification. The underidentification test is a Lagrange Multiplier (LM) test that tests the rank of the matrix to determine whether the equation is identified i.e. whether the excluded instrument is "relevant" or correlated with the endogenous regressor in the first stage selection

equation. Weak identification is when the excluded instrument is weakly correlated with the potentially endogenous regressor. Weak instruments can lead to poor performance of estimators, specifically, estimators might be inconsistent, confidence intervals can be incorrect and the tests for significance of coefficients might lead to wrong conclusions.

Though these tests are designed for a continuous endogenous variable, we proceed with testing by treating our binary endogenous variable as continuous. While this means that the critical values of the tests and the significance levels might be slightly different from those reported here, we note that the CF (endogeneous variable treated as binary) and 2SLS (endogeneous variable treated as continuous) results are nearly identical (see Table 2.5 in Chapter 2).

For the underidentification test, we use the Sanderson and Windermeijer χ^2 Wald statistic as reported using the `ivreg2` routine in Stata 16 (Sanderson and Windmeijer 2016, Baum et al. 2002). Under the null, the equation is underidentified. For our model, the SW χ^2 statistic takes a value of 150,000 with 1 d.f., which has corresponding p-value = 0.00. Hence, there is strong evidence to reject the null hypothesis of underidentification, and we conclude that the excluded instrument is "relevant".

For weak identification, we use the Sanderson and Windermeijer first stage F-statistic which is the F version of the SW χ^2 statistic. It is used as a diagnostic for whether the endogenus regressor is "weakly identified". The F-statistic from the model is compared against the critical values of the Kleibergen-Paap statistic (for cluster robust standard errors) reported in Stock et al. (2005) to determine whether the instruments are weak. The null hypothesis of the test is that the equation is weakly identified and the maximum bias of the IV estimator relative to the bias of OLS is some specified value such as 10%. For a single endogenous regressor with cluster robust standard errors, the Stock-Yogo critical values are 16.38, 8.96, 6.66 and 5.53 for maximum bias of 10%, 15%, 20% and 25%, respectively. If the estimated F-statistic is less than a particular critical value then the interpretation is that the instruments are weak for that level of bias. In our model, the estimated SW F-statistic is equal to 150,000, indicating a maximal bias of (significantly) less than 10%. Hence, this suggests that that there is no evidence to suspect that our models are affected by the problem of weak instruments.

3.8.2 Sensitivity of the instrumental variable

Due to its construction, the instrumental variable is sensitive to low values of $\{C_{gt} \backslash C_{igt}\}$ and $\sum_{t \in T_t} \sum_{c \in \{C_{gt} \backslash C_{igt}\}}$. This leads to long tails in the distribution, though we note that 99% of the values are in the range 0 and 2, with a mean of 1.08 and a median of 1.08. Given the long tails in the IV distribution, we have also tested two alternative characterizations of the IV as robustness:

1. We use a dichotomized version of the instrumental variable following MacCallum et al. (2002), specifically, $BinaryIV = 0$ if $WkAccessibility_{igt} < YrAccessibility_{igt}$ and $BinaryIV =$

1 otherwise. The interpretation is that when $BinaryIV$ is 0, the doctor is less accessible to his patients during that week compared to his yearly average, and if $BinaryIV$ takes a value of 1, the doctor is more accessible to his patients than the yearly average.

2. We censor the instrumental variable at 2, referring to it as the $CappedIV$, and include an additional dummy variable as a control to indicate those observations for which the instrumental variable took a value of greater than 2 (1% of the observations). Similar to the interpretation of $BinaryIV$, when $CappedIV$ is less than 1, the doctor is less accessible to her patients during that week compared to her yearly average, and if the value is greater than 1, the doctor is more accessible to her patients than the yearly average.

Summary statistics for the original IV and for the two variations on the instrumental variable described above are given in Table 3.4. Using $BinaryIV$ and $CappedIV$ in our control function model, the estimated effect sizes take value 13.2% (95% CI: $[12.2\%, 14.1\%]$) and 13.2% (95% CI: $[12.4\%, 14.0\%]$), respectively. Thus, the results are insensitive to the choice of instrument.

Table 3.4 Descriptive statistics for the instrumental variable

	Mean	Median	Min	Max	SD
(1) Instrumental Variable	1.08	1.08	0.00	252.00	0.64
(2) Binary version of the IV ($BinaryIV$)	0.66	1.00	0.00	1.00	0.47
(3) Capped version of the IV ($CappedIV$)	1.07	1.08	0.00	2.00	0.32

3.8.3 IV control

One concern is that there may be omitted variables that threaten the validity of the instrumental variable, particularly if there are unobserved factors that correlate with both the relative accessibility of the regular doctor for other patients (instrument) and the focal patient's revisit interval (dependent variable). For example, if there is a flu outbreak, the focal patient's expected revisit interval might decrease, and at the same time other patients would also find it harder than normal to access their regular provider.

However, such omitted factors should also be expected to affect the revisit interval of the doctor's other patients. We use this insight to construct a control variable to improve the validity of the instrumental variable. Specifically, we add a control that takes the focal patient's expected $ln(RI)$ and adjusts this for the average of $ln(RI)$ of other patients who (i) share the same regular doctor as the focal patient and (ii) visit a doctor in the same week as the focal patient. Intuitively, when the average revisit interval of these other patients change, then the expected $ln(RI)$ of the focal patient should change too.

More specifically, we calculate the control described above for each consultation c_i' that occurred on day t in week w for patient i. To do so, first let $C_{i[t-730,t-1]}$ denote the set of consultations

that occurred between day $t-1$ and $t-730$ for patient i. The time between each consultation c_i made by patient i in that interval is then given by RC_{c_i}. Thus, the past average revisit interval (RI) for consultations that occurred over the two years $(t-730, t-1)$ prior to consultation c_i' is as follows:

$$PastAvgRI_{c_i'} = \frac{\sum_{c_i \in C_{i[t-730,t-1]}} RI_{c_i}}{\left|C_{i[t-730,t-1]}\right|}$$

where $|\cdot|$ specifies the cardinality of the set.

Next, let $g_{c_i'}$ denote the regular doctor associated with patient i at the time of consultation c_i'. Further, let $J_{c_i'w}$ specify the set of all consultations by *other* patients for which the regular doctor is the same as for consultation c_i' (i.e., is $g_{c_i'}$) and that took place in the same week w in which consultation c_i' occurred. Using the same method as described above, we can then calculate the past average revisit interval for those consultations $c_j' \in J_{c_i'w}$, where j specifies the patient associated with consultation c_j'. These revisit intervals are given by $PastAvgRI_{c_j'}$.

Next, we create a multiplier that captures the difference between the current and past revisit intervals for those consultations $c_j' \in J_{c_i'w}$, which is equal to $MultRI_{c_j'} = RI_{c_j'}/PastAvgRI_{c_j'}$. Averaging the multiplier over all consultations $c_j' \in J_{c_i'w}$ we get

$$\overline{MultRI}_{c_i'} = \frac{\sum_{c_j' \in J_{c_i'w}} MultRI_{c_j'}}{|J_{c_i'w}|}.$$

Notice that when $\overline{MultRI}_{c_i'} < 1$ it suggests that revisit intervals were shorter than normal for *other* patients who visited during the same week w in which consultation c_i' took place and who shared the same regular doctor $g_{c_i'}$. Meanwhile, when $\overline{MultRI}_{c_i'} > 1$ it suggests that revisit intervals were longer than normal.

The multiplier thus allows us to account for unobserved factors that are correlated with both the IV and the outcomes. For example, if there is a flu outbreak, we might expect the multiplier to take a value less than 1, reflecting the fact that parents are expected to return faster on average. Using this insight, we can then adjust the expected revisit interval of the focal patient by taking their past revisit interval and multiplying by this multiplier, giving us:

$$ExpectedRI_{c_i'} = PastAvgRI_{c_i'} \times \overline{MultRI}_{c_i'}$$

This variable thus gives us the expected time to the next visit of patient i based on their past revisit interval and the multiplier calculated from other patients.

Adding $ln(ExpectedRI_{c_i'})$ as an additional control variable in our model allows us to account for various unobserved factors that might be correlated with both the availability of the regular doctor for other patients (the instrument) and also the expected revisit interval of the focal patient (the dependent variable).[5] Summary statistics for this control variable are given in Table 3.5.

[5]We use the natural logarithm of the control to match with the units of the dependent variable (natural logarithm of the revisit interval).

Table 3.5 Descriptive statistics for the control for the instrumental variable

	Mean	Median	Min	Max	SD
(1) $ln(ExpectedRI)$	4.08	4.10	-3.88	10.25	0.77

3.9 Confounding: Subsample analysis

As outlined in Section 2.5.3, patient acuity is the most obvious source of omitted variable bias. Specifically, higher acuity patients may be unwilling to wait to see their regular provider, and hence are more likely to see their non-regular doctor. Moreover, these acute patients are more likely to return sooner after the focal consultation for a follow-up visit after the index consultation. This would lead to a shorter revisit interval and could provide an alternative, non-causal explanation for β_i in the OLS model.

One way we address endogeneity concerns is by subsampling the data and comparing the effect sizes associated with seeing the regular doctor across subsamples that indicate higher acuity patients. In selecting subsamples, we aim to identify potentially "acute" consultations for which patients have a higher likelihood of seeing a non-regular doctor and also a shorter revisit interval. We wish to establish that our results are consistent across acute and non-acute consultations, so we perform the following analyses.

Antibiotic prescriptions

We first consider samples based on prescriptions. Specifically, we identify (i) the sample of consultations when no medication was prescribed and (ii) the sample of consultations when a new antibiotic was prescribed. Antibiotic prescriptions are identified using Gulliford et al. (2020). The rationale is that patients who are prescribed a new antibiotic during their consultation are more likely than other patients to have an acute problem. We estimate the same model as in Equation 2.1 (re-written below), using the same dependent variable and the same set of control variables:

$$ln(RI_i) \quad = \quad \beta_0 + \beta_1 RD_i + \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i \,, \tag{3.1}$$

where the vector $\mathbf{X}_i$ specifies the set of controls corresponding to observation i, as defined in Section 2.4.3.4, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is the error term.

If acuity is in fact a major confounder, we expect that the acute consultations would have a smaller value of β_1 in Equation 3.1. The results (reported in Table 3.6) show that the effect size is indeed slightly smaller, suggesting that the effect is less pronounced for acute conditions. However, the effect remains large and significantly positive.

Table 3.6 Coefficient of RD_i across two different samples: the sample of consultations when (i) no medication was prescribed, (ii) when a new antibiotic was prescribed.

Dependent Variable = Natural logarithm of the revisit interval		
Subsample:	No New Prescription	New Antibiotic Prescription
RD_i=1	0.150^{***}	0.114^{***}
	[0.147,0.152]	[0.110,0.119]
Observations	6,310,933	1,354,977

Notes: 95% confidence intervals in square brackets; The total number of observations in this table is not equal to the total sample size of 11,344,065 observations, as there is a category of observations at which a non-antibiotic medicine was prescribed, which we do not include in this analysis.

Prior emergency department visit

Consultations are more likely to be acute if they are preceded by an emergency department (ED) visit. For example, a patient visiting the ED for an acute reason might be asked by the hospital doctors to follow up with a doctor at their primary care practice. Thus, we class those consultations that are preceded by an ED visit in the seven days prior to the consultation as more acute. We compare the effect of seeing the regular doctor on the revisit interval on two different subsamples: (i) when the index consultation was preceded by an ED visit within seven days and (ii) the remainder of the sample, which had no ED visits within seven days before the index consultation.

Again, we re-estimate the same model as in Equation 3.1. If the results are driven by acute consultations only, we would expect that the effect size of the subsample where the consultation was preceded by an ED visit would have a much larger effect size. However, as we report in Table 3.7, the effect size is large and significant in both cases.

Table 3.7 Coefficient of RD_i across two different samples: the sample of consultations that were (i) preceded by an ED visit in the 7 days prior to the consultation, and (ii) those that were not.

Dependent Variable = Natural logarithm of the revisit interval		
Subsample:	ED visit 7 days prior	No ED visit 7 days prior
RD_i=1	0.098^{***}	0.137^{***}
	[0.084,0.112]	[0.136,0.139]
Observations	175,123	11,168,942

Notes: 95% confidence intervals in square brackets

Diabetes Medication Review

Another way to identify non-acute consultations is to focus our attention on diabetes patients, a comorbidity that requires disease management and has well-defined measures to monitor the health of the patient (Leniz and Gulliford 2019, Worrall and Knight 2011). From all the consultations with diabetes patients, we specifically consider those consultations that were coded as "Diabetes review". These consultations are most likely to be a part of disease management and not acute in nature. We compare the coefficient of RD_i across diabetes consultations that are (i) any type of diabetes consultations (ii) diabetes consultations excluding review consultations and (iii) diabetes consultations that are only review consultations. Diabetes consultations are identified using the code list in Payne et al. (2020). Results are shown in Table 3.8 and suggest that the effect is consistent across the samples.

Table 3.8 Coefficient of RD_i across (i) base sample (ii) any type of diabetes consultations (iii) diabetes consultations excluding review consultations and (iv) diabetes consultations that are only review consultations.

Dependent Variable = Natural logarithm of the revisit interval				
Subsample:	Base	Diabetes (Any)	Diabetes (excluding review)	Diabetes (only review)
Appt with Regular doctor	0.156^{***}	0.179^{***}	0.178^{***}	0.180^{***}
	[0.147,0.164]	[0.168,0.191]	[0.166,0.190]	[0.118,0.243]
Practice level controls	Yes	Yes	Yes	Yes
Patient level controls	Yes	Yes	Yes	Yes
Practice FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Observations	11,344,065	1,314,964	1,299,351	15,618

Notes: 95% confidence intervals in square brackets

3.10 Summary statistics for the matched sample

In Table 3.9 we compare the proportions of the control variables for the full sample with those corresponding to the control group and the treatment group. (Specifically, in this table we compare the patient demographics across the different groups. Comparison across other factors (not shown) also suggests that there is no significant different between the groups.) We also report p-values corresponding to Pearson χ^2 tests for differences in the proportions between the control and treatment groups in the matched sample. As can be seen, the proportions remain the same across the three columns of Table 3.9, as desired.

Table 3.9 Comparing the balance of the full sample with the control group and treatment group of the matched sample.

	Full sample (%)	Control group (%)	Treated group (%)
Age band			
18-25yrs	5.58	5.62	5.54
26-30yrs	4.68	4.68	4.67
31-35yrs	5.33	5.35	5.31
36-40yrs	6.16	6.14	6.18
41-15yrs	7.17	7.13	7.21
46-50yrs	7.82	7.78	7.85
51-55yrs	7.94	7.94	7.94
56-60yrs	8.02	8.03	8.01
61-65yrs	8.94	8.94	8.94
66-70yrs	8.60	8.63	8.57
71-75yrs	8.30	8.27	8.32
76-80yrs	7.99	8.01	7.97
81-85yrs	6.62	6.65	6.59
86+yrs	6.84	6.81	6.88
		p-value for Pearson test (χ^2)=0.913	
Comorbidity			
0 comorbidites	24.46	24.50	24.41
1 comorbidity	25.14	25.09	25.20
2 comorbidites	19.15	19.10	19.19
3 comorbidites	13.08	13.10	13.06
4 comorbidites	8.26	8.27	8.24
≥ 5 comorbidites	9.92	9.93	9.91
		p-value for Pearson test (χ^2)=0.850	
Mental Health			
No	72.80	72.73	72.86
Yes	27.20	27.27	27.14
		p-value for Pearson test (χ^2)=0.328	
Gender			
Male	37.56	37.62	37.50
Female	62.44	62.38	62.50
		p-value for Pearson test (χ^2)=0.406	
Index of Multiple Deprivation			
1	22.84	22.87	22.81
2	21.58	21.52	21.64
3	20.82	20.82	20.82
4	18.43	18.42	18.54
5	16.34	16.38	16.30
		p-value for Pearson test (χ^2)=0.813	
Prescriptions			
0	23.07	23.16	22.99
1	12.71	12.63	12.79
2	9.91	9.87	9.95
3	8.37	8.36	8.38
4-5	13.94	13.96	13.92
6-7	11.03	11.03	11.03
8-9	7.95	7.96	7.94
10-12	7.17	7.21	7.13
13-15	3.45	3.42	3.48
16+	2.40	2.40	2.40
		p-value for Pearson test (χ^2)=0.581	

Notes: The null hypothesis for the Pearson χ^2 test is that there is no difference between the two proportions. In each case, we find that we cannot reject the null at the 5% significance level.

3.11 Results for the duration of consultations

Table 3.10 summarises the results of all the models introduced in Section 2.5 but using the length of consultation as the dependent variable. The results show that the regular doctors spend less time on average with their patients than transactional providers. Hence, providing care continuity will not come at a cost of reducing daily clinical throughput in primary care practices. However, the effect size is not large enough to be actionable. Physicians typically work in four-hour shifts and will be unable to accommodate an additional patient.

Table 3.10 Coefficient of RD_i across all modeling techniques.

Dependent variable = Length of the consultation (minutes)								
	Sample	Model	Coefficient	Std. Error	t-statistic	P> $\|t\|$	95% CI	
1a	consultations	OLS	-0.37	0.01	-66.0	0.00	-0.38	-0.36
1b	consultations	CF-IV	-0.15	0.03	-5.01	0.00	-0.20	-0.09
1c	consultations	IV-2SLS	-0.11	0.03	-3.67	0.00	-0.17	-0.05
2a	patients	PSM	-0.32	0.02	-13.86	0.00	-0.37	-0.27
2b	patients	PSM OLS	-0.32	0.02	-14.47	0.00	-0.37	-0.28
3a	patients	PSM 0.25<p<0.75	-0.31	0.02	-13.20	0.00	-0.36	-0.27
3b	patients	PSM OLS 0.25<p<0.75	-0.32	0.02	-13.79	0.00	-0.36	-0.27
4a	patients	PSM 0.33<p<0.67	-0.31	0.03	-12.02	0.00	-0.36	-0.26
4b	patients	PSM OLS 0.33<p<0.67	-0.31	0.02	-12.57	0.00	-0.36	-0.26
5a	patients	PSM 0.4<p<0.6	-0.32	0.02	-13.86	0.00	-0.37	-0.27
5b	patients	PSM OLS 0.4<p<0.6	-0.32	0.02	-14.47	0.00	-0.37	-0.28

Notes: Standard errors are clustered at the patient level for 1a and at the pair level for 2a-5b; This table reports the coefficient estimates of care continuity for the taxonomy of models specified in this section. OLS regression refers to the ordinary least squares regression as specified in Equation 2.1. CF-IV refers to the control function approach specified in Section 2.5.3.1 and 2SLS refers to the two-stage least squares method using the same IV as used for the CF-IV approach. PSM corresponds to the matching-based effect estimation where we report differences in averages of ln(RI) between the control and treated groups, using nearest neighbor matching without replacement, and a caliper of 0.001 (Section 2.5.4.2); PSM OLS corresponds to an OLS regression on the matched sample that includes the covariates (Section 2.5.4.2); Models 3a-5b correspond to the minimum bias estimators that use subsamples determined by the probabilities (p) (Section 2.5.4.3). $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.10$.

3.12 Moderation results

In Table 3.11, we report coefficient estimates corresponding to the moderation models described in Chapter 2. The interpretation of the coefficients in this table is described in the notes under the table. Additionally, Figure 3.8 shows a graphical representation of the estimates of the average marginal effects based on the moderation results corresponding to Table 2.6 in Chapter 2.

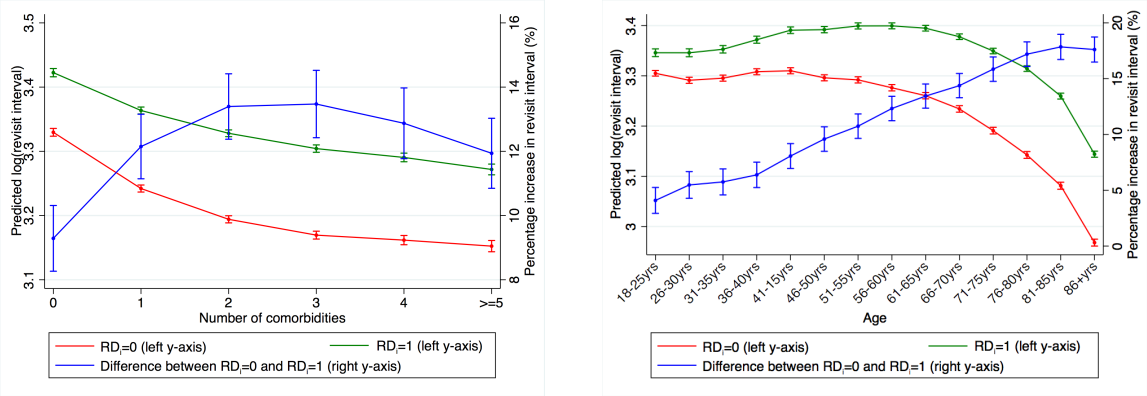

Figure 3.8 Percent increase in ln(Revisit Interval) when the patient sees a transactional provider ($RD_i = 0$) compared to when the patient sees her regular doctor ($RD_i = 1$), (left) for patients with multiple chronic diseases (H2a) and (right) across different age groups (H2b)

Table 3.11 Moderating effects of age, comorbidities and mental health, using the control function approach

CF: Dependent variable = Natural logarithm of the revisit interval				
	$RD_i = 0$		Interaction	
	Coefficient	95% CI	Coefficient	95% CI
Baseline ($RD_i = 1$)	–	–	0.011	[-0.001,0.023]
26-30yrs	-0.014***	[-0.019,-0.008]	0.010*	[0.001,0.020]
31-35yrs	-0.010***	[-0.015,-0.004]	0.018***	[0.009,0.027]
36-40yrs	0.001	[-0.004,0.006]	0.030***	[0.021,0.039]
41-15yrs	0.000	[-0.005,0.006]	0.040***	[0.032,0.049]
46-50yrs	-0.006*	[-0.011,-0.000]	0.052***	[0.043,0.060]
51-55yrs	-0.013***	[-0.019,-0.008]	0.065***	[0.057,0.074]
56-60yrs	-0.029***	[-0.035,-0.023]	0.083***	[0.075,0.091]
61-65yrs	-0.042***	[-0.048,-0.036]	0.090***	[0.082,0.098]
66-70yrs	-0.074***	[-0.080,-0.068]	0.103***	[0.094,0.111]
71-75yrs	-0.114***	[-0.120,-0.108]	0.115***	[0.107,0.124]
76-80yrs	-0.162***	[-0.168,-0.155]	0.128***	[0.120,0.137]
81-85yrs	-0.229***	[-0.236,-0.222]	0.142***	[0.133,0.151]
86+yrs	-0.337***	[-0.344,-0.331]	0.132***	[0.123,0.141]
1 comorbidity	-0.088***	[-0.091,-0.084]	0.029***	[0.024,0.033]
2 comorbidities	-0.136***	[-0.140,-0.131]	0.041***	[0.036,0.046]
3 comorbidities	-0.160***	[-0.167,-0.154]	0.042***	[0.036,0.048]
4 comorbidities	-0.168***	[-0.176,-0.160]	0.036***	[0.029,0.043]
≥ 5 comorbidities	-0.177***	[-0.187,-0.166]	0.027***	[0.020,0.033]
Mental health condition	-0.015***	[-0.018,-0.011]	0.025***	[0.021,0.028]

Notes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$; 95% confidence intervals in square brackets, with standard errors clustered at the patient level; 'Baseline' row gives the effect of seeing the regular doctor for an 18-25 year old with no comorbidities and no mental health condition; The '$RD_i = 0$' column specifies the effect of age, comorbidity and mental health for a patient who did not see their regular doctor; The effect of age, comorbidity and mental health for a patient who saw their regular doctor can be determined by taking the baseline effect, and adding to this the sum of the coefficients from the '$RD_i = 0$' and 'Interaction' columns.

3.12.1 Communicating results with doctors

Since the comorbidity scores from primary care data are tedious and involved to calculate, it is easier for doctors to categorize patients by their medication pattern. Hence, we repeat our moderation analysis with the number of different active repeat prescriptions the patient is prescribed within a 6-month window preceding the consultation, in order to provide results that can be more easily communicated to doctors. Similar to the comorbidities, we find that for each category (as described in 2.3 of Chapter 2), the revisit interval is longer when the patient sees his regular doctor (green line in Figure 3.9), and the differential effect increases linearly till 3 prescriptions, after which the effect remains stable (blue line in Figure 3.9).

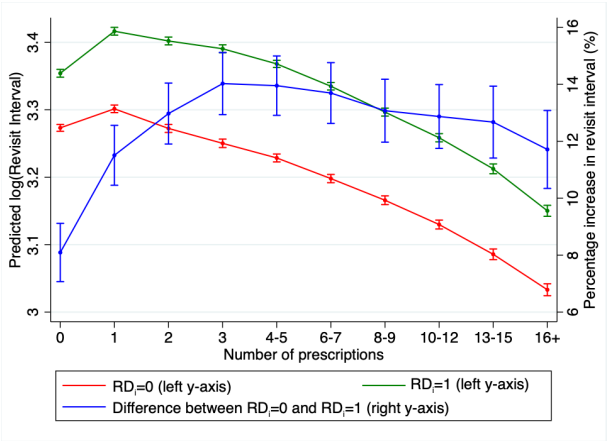

Figure 3.9 Percent increase in ln(Revisit Interval) when the patient sees a transactional provider ($RD_i = 0$) compared to when the patient sees her regular doctor ($RD_i = 1$), for patients on multiple prescriptions

3.13 Relationship between age and comorbidity

Figure 3.10 combines the effect of providing care continuity to patients in different age groups who have either 0, 1 or 2 comorbidities. This is estimated by including the following interaction. The largest productivity-enhancing effect comes from providing care continuity to patients 86+ years of age with one comorbidity (21.0%; 95% CI: $[19.2\%, 22.8\%]$). The productivity gain from providing care continuity to younger patients with no comorbidities compared to younger patients with two comorbidities is much higher than for older patients; for older patients, in terms of productivity gains from providing care continuity, it does not make a significant difference whether the patient has 0, 1 or 2 comorbidities.

3.13.1 Continuity and healthy patients

We run additional analysis to communicate that transactional medicine has a place in primary care for younger, healthy patients.

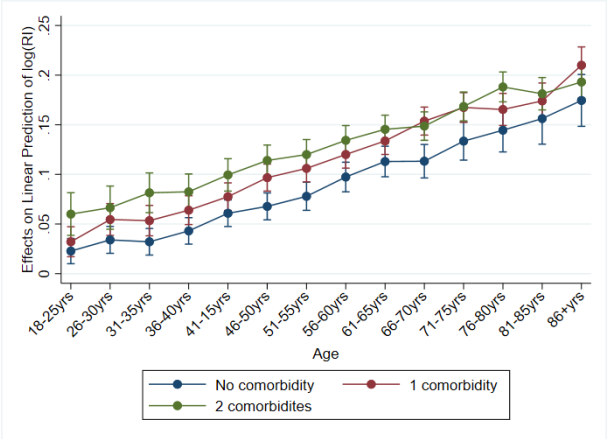

Figure 3.10 Differential effect of providing care continuity on RI for patients across different age groups with 0, 1 or 2 comorbidities

To estimate the effect of continuity of care on revisit intervals for healthy patients, we re-estimate the CF model (Equations Equation 2.2 and Equation 2.3) but now also include an interaction term between the main independent variable RD_i and an indicator variable which takes a value of 1 for patients below the age of 60 who have no comorbidities and no reported mental health conditions.

Table 3.12 Results from the interaction model that include include an interaction term between the main independent variable RD_i and an indicator variable which takes a value of 1 for patients below the age of 60 who have no comorbidities and no reported mental health conditions.

Dependent Variable = Natural logarithm of the revisit interval		
Variable	Coefficient	95% CI
$RD_i = 1$	0.17	[0.17, 0.17]
Healthy patient indicator	0.13	[0.12, 0.13]
$RD_i = 1$*Healthy patient indicator	-0.09	[-0.09, -0.08]

The interpretation for the model from Table 3.12 is as follows:

- Moving from the omitted category (old or unhealthy, no-continuity) to (young and healthy, no-continuity) increases the revisit interval by 13% (effect of young and healthy for transactional provider)
- Moving from the omitted category (old or unhealthy, no-continuity) to (young and healthy, continuity) increases the revisit interval by 21%, and therefore (because of (i)), moving from (young and healthy, no-continuity) to (young and healthy, continuity) increases the interval by 7% (effect of continuity for young and healthy)
- Moving from the omitted category (old or unhealthy, no-continuity) to (old or unhealthy,

continuity) increases the revisit interval by 17% (effect of continuity for the old or unhealthy)

- Moving from the omitted category (unhealthy, continuity) to (healthy, continuity) therefore increases the revisit interval by 4% (=21%-17%) (effect of young and healthy for regular provider)

We summarize these results into a matrix by segmenting the service provided (regular vs. transactional) and the type of patient (young and healthy vs. old or unhealthy). We find that the regular provider is able to keep the old or unhealthy away from the practice almost as long as the transactional provider does for the young and healthy.

Predicted ln(Revisit Interval)	Young and healthy	Old or unhealthy
Transactional service	3.32	3.19
95% CI	[3.31, 3.32]	[3.19, 3.19]
Regular provider	3.40	3.36
95% CI	[3.39, 3.40]	[3.36, 3.36]

3.14 Targeting continuity of care

3.14.1 Are primary care practices offering targeted continuity?

To summarize our analysis, for a primary care practice manager, the key decision is: shall we pool doctor time and offer it on a first-come-first-served basis to improve fast access or shall we reserve some of that time to ensure continuity of care and improve relationships between doctors and patients? To answer this question, we look at how continuity of care affect the productivity of the clinical workforce and find that offering continuity increases revisit intervals by an estimated 13% without requiring longer consultations; the improvement is about 17% for patients above 60 or with chronic illnesses. Hence, we recommend that practices should make sure that older and more complex patients are receiving more continuity. This is an important insight because as we can see from Figure 3.11 that continuity is declining at the same rate in the old and young and hence practices are not prioritizing continuity in this manner.

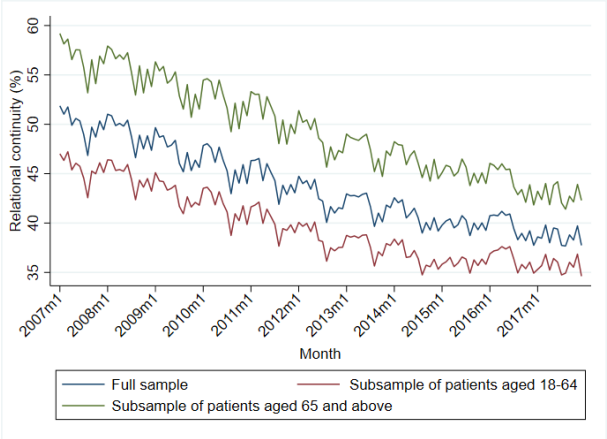

Figure 3.11 Continuity of care over time for different segments of patient age

Calculation of the scoring index for targeting continuity of care

Our results suggest that practices should increase their continuity offering to improve productivity. However, we do not recommend that practices offer increased continuity on a random basis, but take into account our results to design the reallocation of continuity to patients who benefit from it the most. Using the insights from Hypotheses 2, 3 and 4, we propose a scoring system that can be used by practice managers to prioritize those patients for whom care continuity has the most productivity-enhancing effect and target them for more relational services. Given the current lack of tools to monitor continuity or to measure how well practices have been providing continuity of care to patients who benefit from it, we believe this methodology is an important practical contribution in its own right (Palmer et al. 2018, Hill and Freeman 2011).

The proposed scoring system ranks consultations by the estimated number of days gained if the consultation is offered by the patient's regular doctor as compared to another doctor. To determine the revisit interval days gained, we begin by estimating an interaction model in which we include interactions between the binary indicator for a consultation with the patient's regular doctor (i.e., RD_i) and every other independent variable, using the same method as in Equation Equation 2.2.[6] For each consultation i, we then use this model to predict the revisit interval under two scenarios: (i) assuming the patient receives care continuity ($RD_i = 1$), which we denote $ln(\hat{RI}_1)_i$, and (ii) assuming the patient does not receive care continuity ($RD_i = 0$), which we denote $ln(\hat{RI}_0)_i$. In order to calculate the productivity gain in number of days rather than in logarithmic units, we transform $ln(\hat{RI}_0)_i$ and $ln(\hat{RI}_1)_i$ using Duan's smearing transformation to get $\hat{RI}_{0i}$ and $\hat{RI}_{1i}$ respectively, in days.[7] For each patient-consultation, we

[6] This model gives us the best in-sample fit for the predicted revisit interval.

[7] In a model $ln(y_i) = \mathbf{x}\beta + \epsilon_i$, if the errors are normally distributed as $N(0, \sigma_\epsilon^2)$, the estimates can be recovered as follows: $E[y|\mathbf{x}] = exp(\mathbf{x}\hat{\beta} + 0.5\hat{\sigma}_\epsilon^2)$ where $\hat{\sigma}$ is the usual unbiased estimator of σ. But if the errors are not normally distributed, as it is in our case according to the Jarque-Bera test for normality, then Duan's smearing estimator is used. These estimates are recovered as: $E[y|\mathbf{x}] = exp(\mathbf{x}\hat{\beta}) * (N^{-1} \sum exp(\epsilon_i))$ where ϵ_i is the i^{th} residual from the model and $N^{-1} \sum_i exp(\epsilon)$ is the smearing factor (Duan 1983).

define $days\_gained_i = \hat{RI}_{1i} - \hat{RI}_{0i}$, which is the expected days gained or "score" in days from providing care continuity for consultation i.

Once we have calculated the scores, we can calculate the maximal productivity advantage that a practice would have gained from care continuity if it had used a specific proportion of its overall consultations as its care continuity consultation budget. The maximal number $total\_days\_gained_p$ that practice p can gain, relative to offering no continuity of care, is the sum of consultation-level $days\_gained_i$ achieved by allocating regular doctors to consultations in rank order of $days\_gained_i$, from highest to lowest, until the practice runs out of its continuity of care consultation budget.

To apply this to our data, let C_p denote the set of recorded consultations i that took place in practice p, and let N_p be the total number of these consultations. Then for all consultations $i_p \in C_p$, we can sort the $days\_gained_{i_p}$ in descending order and denote this $\text{sort}(days\_gained_{i_p})^{(a)}$ where a corresponds to the rank once the consultations are sorted using this sorting scheme. Next, let $x_p \in [0\%, 100\%]$ specify the percent of continuity provided in practice p. Then the total days gained by offering $x_p\%$ continuity to the most productivity-enhancing patients at practice p can be calculated as

$$total\_days\_gained_p^{[x_p]} = \sum_{a, a \in [0, N_p x_p]} \text{sort}(days\_gained_{i_p})^{(a)}.$$

Change in practice demand by re-allocating continuity

First, note that an extension of the average revisit interval by a factor r leads to a demand reduction of $r/(1+r)$.[8] Therefore, to estimate the effect of continuity on practice demand we can equivalently focus on estimating the impact on the revisit intervals (i.e., estimating the value of r).

To identify r, we start by observing that $\sum_{i_p \in C_p} \hat{RI}_{0i_p}$ gives the sum over the predicted revisit intervals of all consultations C_p that occurred at practice p under the assumption that *no* patient saw their regular doctor. Based on the results in Chapter 2, we should expect intuitively that as more patients see their regular doctor these revisit intervals should increase. For example, $\sum_{i_p \in C_p} \hat{RI}_{0i_p} + total\_days\_gained_p^{[x_p]}$ gives the sum of predicted revisit intervals assuming instead that the $x_p\%$ most productivity-enhancing patients at practice p saw their regular doctor, while all other patients saw a transactional provider. Meanwhile, $\sum_{i_p \in C_p} RI_{i_p}$ (i.e., the sum of the revisit intervals as given in the raw data) gives the sum of the revisit intervals under the status quo, i.e., assuming no change in the percentage of consultations between patients and their regular doctor and no change in the allocation of continuity to consultations. Combining these

[8] Assuming that each patient i has a demand for x_i consultations over a period $[0, T]$, an extension of all revisit intervals by a factor r extends this demand to x_i consultations over a period $[0, (1+r)T]$ or, equivalently, to a demand of $x_i/(1+r)$ consultations over the period $[0, T]$. Demand is therefore reduced by a factor $r/(1+r)$ to the fraction $1/(1+r)$ of the original demand.

two measures, we can estimate the percentage increase in the revisit interval if continuity at a practice were retained at the same level, which we denote x_p^*, but if continuity were instead re-allocated from those patients who actually received it to those patients who stood to benefit from it the most. This is given by the following expression:

$$\frac{\sum_{i_p \in C_p} \hat{RI}_{0i_p} + total\_days\_gained_p^{[x_p^*]}}{\sum_{i_p \in C_p} RI_{i_p}}$$

Subtracting 1 from this expression gives us the factor r, i.e., the increase in the revisit interval by better targeting care continuity. Observe that we should expect numerator to be greater than the denominator (i.e., $r > 0$), since in both the numerator and denominator we offer continuity to the same proportion of consultations ($x_p^*\%$) but the consultations in the numerator are allocated to the most productivity-enhancing patients.

Using r as calculated above, we can then estimate for each practice the reduction in demand that would have been expected from better allocating continuity while keeping continuity-levels unchanged. The productivity-gains for each practice are plotted in Figure 2.2 in Chapter 2.

Change in system demand by re-allocating and increasing continuity

While in the above analysis we have focused on re-allocating continuity within a practice, we can also ask what would happen if practices increased the level of continuity provided *above* their current level x_p^*. To do this, we start by estimating the total reduction in system demand if all practices had re-allocated continuity to the most productivity-enhancing patients. This is given by the expression:

$$\frac{\sum_p \left(\sum_{i_p \in C_p} \hat{RI}_{0i_p} + total\_days\_gained_p^{[x_p^*]} \right)}{\sum_p \sum_{i_p \in C_p} RI_{i_p}}$$

This is the baseline result if all practices kept continuity at their current levels x_p^*. Now let P_x denote the set of practices with current continuity levels below $x\%$, i.e., with $x_p^* < x$. Then we ask what would have happened to demand if all practices with $x_p^* < x$ had increased continuity levels to x *and* had re-allocated continuity to the $x\%$ most productivity-enhancing patients. (Note that we will assume that all practices with $x_p^* \geq x$ keep continuity at the current level x_p^* but also improve targeting of continuity). For a given minimum continuity level x, the total reduction in system demand is then given by the expression

$$\frac{\sum_{p \in P_x} \left(\sum_{i_p \in C_p} \hat{RI}_{0i_p} + total\_days\_gained_p^{[x]} \right) + \sum_{p \notin P_x} \left(\sum_{i_p \in C_p} \hat{RI}_{0i_p} + total\_days\_gained_p^{[x_p^*]} \right)}{\sum_p \sum_{i_p \in C_p} RI_{i_p}}$$

We can estimate this for all values $x \in [0\%, 100\%]$. This is shown in Figure 2.3 in the Chapter 2.

3.15 Effect of continuity of care on healthcare resource consumption and downstream resources

There are various studies which look at the effect of continuity of care on healthcare utilization, such as prescription rates, referral rates and ED visits. To complement this line of research, we investigate the effect of seeing the patient's regular provider on such health care resources. This analysis also helps us understand what it is that the continuity of care doctor does differently compared to a transactional provider, for example, whether the regular provider or transactional provider is more likely to make decisions that would increase the healthcare consumption of the patient. We look at various aspects of the consultation, specifically activities that take place (1)"during" the focal consultation and activities that take place in the (2) "after" the focal consultation.

1. During the consultation: As outlined in Hypothesis 2,3 and 4, the patient's regular provider may be able to make more informed clinical decisions as he possesses superior information about the patient's overall needs, which can ultimately extend the time to subsequent appointment. A transactional doctor may be more hesitant to prescribe a medication to a patient he is not familiar with, and therefore may order extra tests to be sure. But, on the other hand, a transactional doctor may also simply prescribe medication to treat the symptoms without understanding the root cause of the problem, knowing that future workload will be shared by other doctors at the practice. Similarly, a regular doctor might also practice expectant management with the patient he is familiar with or be more confident in prescribing and referring a patient to a specialist. Whether the doctor who is more familiar with the patient spends more or less time with the patient, prescribes more or less medication or exercises better judgement with regards to referrals can be hypothesized, but remains an empirical question.

- Consultation Length: Several studies are concerned with the consultation length or visit duration as the physician's time is an important resource and can be affected by the doctor's knowledge of the patient (Dugdale et al. 1999, Irving et al. 2017, Hjortdahl and Borchgrevink 1991). On the one hand, the regular doctor might wish to spend more time with the patient to treat the patient thoroughly (Jeffers and Baker 2016, Mercer et al. 2007). On the other hand, with increased knowledge of the patient, the regular doctor can save time during the consultation (Hjortdahl and Borchgrevink 1991, Hill and Freeman 2011, Rosen et al. 2020). Given that this is UK data, we do not expect much of a difference in the duration of consultations with a patient's regular doctor compared to the transactional doctor, since 10 minutes has become the standard consultation length for primary care visits. But, we do find that (Section 3.11) on average, the consultation length is slightly lower for consultations with the patient's regular doctor, indicating that regular doctor allocates his time better but not at the expense of increased duration or congestion.

- Referrals: There isn't a clear consensus in the literature regarding the effect of continuity of care on referral rates. A Dutch study finds that increased continuity of care decreases referral to specialist care, most notably for referrals to paediatrics (Olthof et al. 2019). Regular doctors are accountable for the health of their patient and bear the future workload of the patient's visit themselves, so they might be more keen to wait and see before they refer the patients, while transactional doctors might be more likely to use referrals as an alternative "pass the buck" route. Hjortdahl and Borchgrevink (1991), on the other hand, find that patients had a statistically significant, twofold increased chance of being referred if the doctors knew their medical history. In line with this, we find some evidence for an increase in referrals when the patient sees his regular doctor, though the effect size is very small.

- Prescriptions: A transactional doctor who is unfamiliar with the patient might be more reluctant to prescribe certain types of drugs. On the other hand, a transactional doctor may simply prescribe a drug as he doesn't have the incentive to understand the root cause of the problem, but to alleviate symptoms. At the same time, a continuity of care doctor who is familiar with the patient's condition may ask the patient to wait and see before prescribing any drugs or might even be more willing to over-treat Bobroske et al. (2021). Therefore, whether the regular provider of the patient has a higher prescribing rate than a transactional provider is an empirical question. We find evidence that the regular provider is less likely to prescribe medication that a transactional provider, though the effect size is very small.

To investigate the effect of seeing the regular doctor on referrals and prescriptions, we estimate the following linear probability modes:

$$ResourceUse \quad = \quad \beta_0 + \beta_1 RD_i + \boldsymbol{\beta}\mathbf{X} + \epsilon\,. \tag{3.2}$$

where $ResourceUse$ corresponds to (i) referrals (ii) any prescription and (iii) antibiotic prescriptions, vector $\mathbf{X}$ is the set of controls as defined in Section 2.4.3.4, $\epsilon_{pm} \sim \mathcal{N}(0, \sigma^2)$ is the idiosyncratic error term and standard errors are clustered at the patient level to account for autocorrelation within the same patient. We find from Table 3.13 that when a patient sees his regular doctor, the probability of a referral is higher than when the patient sees a transactional provider, though the effect is very small. The results from Table 3.14 suggest that when a patient sees his regular doctor, the probability of a prescription (general and antibiotic) is lower than when the patient sees a transactional provider.

	Referral (0/1)
Appt with Regular Doctor=1	0.005***
	[0.005,0.005]
All controls	Yes
Observations	12929329

95% confidence intervals in brackets

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3.13 Effect of seeing the patient's regular provider on the referral rate

	New prescription (0/1)	New Antibiotic (0/1)
Appt with Regular Doctor=1	-0.011***	-0.045***
	[-0.012,-0.011]	[-0.045,-0.045]
All controls	Yes	Yes
Observations	11,344,065	11,344,065

95% confidence intervals in brackets

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3.14 Effect of seeing the patient's regular provider on the prescribing rate

2. Continuity of care and downstream activities after the primary care consultations

Depending on the quality of consultation or priorities of the practice, downstream activities might be affected as demand can shift to alternative channels. A lack of familiarity with the patient may increase uncertainty when a non-continuity of care doctor is deciding whether to self-treat the patient or else refer them for further investigation. In such cases, doctors are more likely to err on the side of caution and refer the patient to a downstream provider rather than risk treating the patient themselves (Freeman et al. 2020). Second, since the non-continuity of care doctor is not the patients' primary doctor, they lack accountability for the patient and may be more likely to "pass the buck" to someone else.

There is extensive literature on the effect of continuity of care on downstream activities as outlined in Section 2.2.2 of Chapter 2. These studies collectively suggest that offering continuity to patients reduces the utilization of the A&E.

To study the effect of seeing the regular doctor on A&E utilization in our setting, we aggregate our dataset into a panel with two levels, practice and year. The main independent variable is the average RD_i across the practice-year, RD_{pt}, where p corresponds to the practice and t corresponds to the year. The main two dependent variables are (i) $ln(AEadmissions)_{pt}$: logarithm of the number of A&E visits from patients in practice p during year t and (ii) $ln(AEBedDays)_{pt}$: logarithm of the number of bed days resulting from A&E admissions from patients in practice p during year t. We also include various control variables at the practice-year level, such as, the average service duration, average age of the patients, average gender breakdown of the patients and the average consultation frequency of the patients.

We estimate four fixed-effects regression models as follows:

$$ln(AEadmissions)_{pt} = \alpha_p + \beta_1 RD_{pt} + \mathbf{X}_{pt} + \epsilon_{pt} \tag{3.3}$$

$$ln(AEadmissions)_{pt} = \alpha_p + \beta_1 RD_{pt-1} + \mathbf{X}_{pt-1} + \epsilon_{pt} \tag{3.4}$$

$$ln(AEBedDays)_{pt} = \alpha_p + \beta_1 RD_{pt} + \mathbf{X}_{pt} + \epsilon_{pt} \tag{3.5}$$

$$ln(AEBedDays)_{pt} = \alpha_p + \beta_1 RD_{pt-1} + \mathbf{X}_{pt-1} + \epsilon_{pt} \tag{3.6}$$

Practice-specific intercepts, α_p, capture unobserved time-invariant heterogeneity across practices. Meanwhile, the vector $\mathbf{X}_{pm}$ contains the set of control variables and $\epsilon_{pm} \sim \mathcal{N}(0, \sigma^2)$ is the idiosyncratic error term. Standard errors are clustered at the practice level to account for autocorrelation within the same practice.

Equations 3.3 and 3.5 include the independent variable at the same time period as the dependent variable, whereas Equations 3.4 and 3.6 include the lagged level of the independent variable. The interpretation of the two sets of equations are different. In equations 3.3 and 3.5, we look

at a short-term "substitution effect" where the doctor does not avoid a health deterioration that may result in hospitalization, but picks it up early enough and deals with it instead of the hospital. In equations 3.3 and 3.5, we hypothesise a long-term "population health" effect, i.e. the doctor keeps the patient healthier and the patient may therefore not have an episode that requires hospitalization. We do not expect a large effect for the latter.

We report the results for the fixed effects models in Tables 3.15 and 3.16. For both tables, Column 1 and 2 contain results for the level independent variable without and with controls, respectively, and Column 3 and 4 contain results for the lagged independent variable without and with controls, respectively. We find that for the long-term effects (Column 3 and 4 in Tables 3.15 and 3.16), the effect is not significant. For the short-term effects, (Column 1 and 2 in Tables 3.15 and 3.16), the effect sizes are large and in the direction we would expect – higher continuity at the practice reduces the number of emergency visits and emergency bed days, but the statistical significance is very low as we only exploit within-practice variation.

An alternative method to conduct this analysis would be to choose an appropriate cohort of patients and track their A&E attendance over time, as a function of continuity. I plan to carry this out as part of my future research.

Dependent variable = Log(A&E admissions)								
	Level		Level		Lag		Lag	
RD_{pt}	-0.226*	(0.105)	-0.153	(0.109)				
$AvgServiceDuration_{pt}$			-0.004	(0.006)				
$AvgAge_{pt}$			-0.021**	(0.007)				
$AvgMale_{pt}$			0.060	(0.241)				
$AvgPrescriptionRate_{pt}$			0.311+	(0.176)				
$AvgConsFrequency_{pt}$			-0.008+	(0.005)				
RD_{pt-1}					-0.133	(0.097)	-0.093	(0.103)
$L.AvgServiceDuration_{pt-1}$							-0.005	(0.005)
$L.AvgAge_{pt-1}$							-0.014*	(0.006)
$L.AvgMale_{pt-1}$							0.253	(0.230)
$L.AvgPrescriptionRate_{pt-1}$							0.245+	(0.145)
$L.AvgConsFrequency_{pt-1}$							-0.009+	(0.005)
Time Trend	0.084***	(0.005)	0.093***	(0.006)	0.059***	(0.004)	0.065***	(0.005)
Constant	-161.251***	(9.420)	-178.363***	(11.190)	-111.499***	(8.808)	-122.675***	(9.928)
Practice FE	Yes		Yes		Yes		Yes	
Observations	3138		3138		2765		2765	
R^2	0.368		0.379		0.275		0.286	

Standard errors in parentheses

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3.15 Coefficient of RD_{pt} corresponding to Equations 3.3 and 3.5 without and with control variables

Dependent variable = Log(emergency bed days)								
	level		level		lag		lag	
RD_{pt}	-0.165*	(0.069)	-0.154*	(0.067)				
$AvgServiceDuration_{pt}$			0.001	(0.002)				
$AvgAge_{pt}$			-0.001	(0.003)				
$AvgMale_{pt}$			-0.151	(0.230)				
$AvgPrescriptionRate_{pt}$			0.096	(0.079)				
$AvgConsFrequency_{pt}$			0.000	(0.003)				
RD_{pt-1}					-0.113	(0.068)	-0.108	(0.066)
$AvgServiceDuration_{pt-1}$							-0.002	(0.002)
$AvgAge_{pt-1}$							-0.001	(0.003)
$AvgMale_{pt-1}$							-0.066	(0.190)
$AvgPrescriptionRate_{pt-1}$							0.046	(0.083)
$AvgConsFrequency_{pt-1}$							-0.001	(0.003)
Time Trend	0.003^{+}	(0.002)	0.004*	(0.002)	-0.000	(0.002)	0.001	(0.002)
Constant	1.973	(3.653)	-0.128	(4.270)	8.993*	(4.147)	7.572	(4.748)
Practice FE	Yes		Yes		Yes		Yes	
Observations	3138		3138		2759		2759	
R^2	0.016		0.018		0.003		0.005	

Standard errors in parentheses

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3.16 Coefficient of RD_{pt} corresponding to Equations 3.5 and 3.6 without and with control variables

3.16 Continuity of care and health inequalities

Health inequality is the avoidable, unfair and systematic difference in health between groups of people emanating from differences in socio-economic status, and can manifest itself in many ways (Williams et al. 2020). Health inequalities are prevalent in life expectancy, in long term health conditions, mental ill health, avoidable mortality, and for our context, access to health services. In England, social, economic and environmental conditions that shape the risk factors contributing to health inequalities include smoking, alcohol consumption, poor diet, physician inactivity, all of which are concentrated in the most disadvantaged groups. Specifically, income, gender and ethnicity are key factors driving occupation, access to healthy food and access to green space. More deprived areas face a higher disease burden from a more complex patient population, and at the same time these areas are staffed with fewer doctors per head (Williams et al. 2020). Patients in poorer areas have reported difficulty in getting appointments with the doctors they prefer, and have reported worse experience with primary care than their richer counterparts (Nuffield Trust 2018). Moreover, the study found that these patients are also more likely to go to the hospital in an emergency, indicating that planned care might be poor (Nuffield Trust 2018). In another study, the authors found that the inequality gap was widening even further. By 2018, the percentage of respondents reporting a very good or fairly good experience

of making an appointment with a doctor in more deprived areas reduced significantly more than it did in less deprived areas (The Nuffield Trust 2020). To make the situation worse, another study found that practices serving a more deprived population received 7% less funding per adjusted registered patient compared to less deprived areas (Fisher et al. 2020). Suggestions to address health inequalities include hiring more doctors and reviewing funding allocations both of which are costly and might not be feasible without substantive changes in the policy and budget.

Based on our results, we suggest that providing continuity to a targeted sub-segment of the practice population will reduce demand, provide better access to health services to those who need it the most and address the financial shortfall issues of such practices, thereby addressing some of the concerns of unequal access to care.

Specifically, we find that for each deprivation level (Figure 3.12), the revisit interval is shorter as deprivation becomes worse, but if the patient sees his regular doctor, the revisit interval increases, for each deprivation level (green line). The increase is highest for patients at level 2 of the deprivation index (blue line) (15.9%; 95% CI:[15.5%, 16.2%]). Even though the increase is lowest for patients at the highest deprivation level (12.7%; 95% CI:12.3%, 13.1%]), the effect is large and economically significant. Hence, in a capitation based system such as this, where studies have found that the patient adjusted capitation price is inadequate for more deprived areas that already suffer from under-staffing, providing and prioritizing continuity for the most deprived patients will be highly beneficial in curbing demand to manage the disease burden of the patient population. For practices that serve a high deprivation population where nurses are more available than doctors, these practices could also consider a team based continuity approach that involves both doctors and nurses.

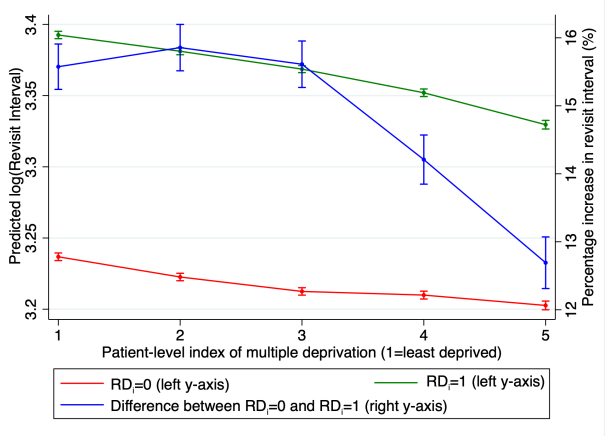

Figure 3.12 Percent increase in ln(Revisit Interval) when the patient sees a transactional provider($RD_i = 0$) compared to when the patient sees her regular doctor($RD_i = 1$), for patients at different deprivation levels

We also look at the differences in providing continuity between rural and urban areas. Though both rural and urban areas have their own unique challenges, in rural areas only 19% of the

patients live within a 20 minute walk from their primary care doctor compared to urban areas, where 94% of the patients living within a 20 minute walk (Todd et al. 2014). Moreover, the ongoing doctor workforce crisis is affecting rural areas disproportionately (Verma et al. 2016). We find that in both rural and urban areas, if the patient sees his regular doctor, the revisit interval is higher (green line in 3.13), though the percentage increase is significantly higher for rural areas compared to urban areas (blue line), suggesting that providing continuity in such practices can be beneficial in combating increasing demand with a shrinking workforce.

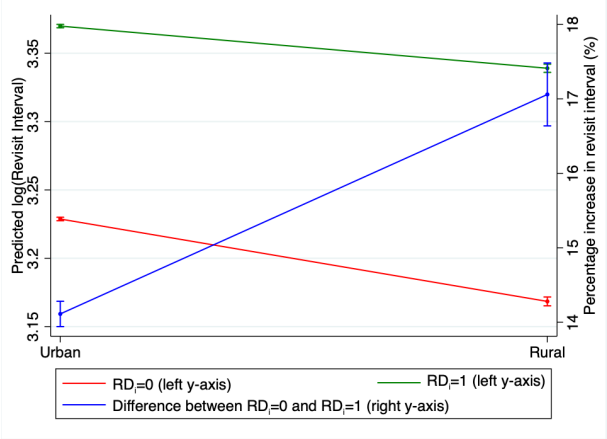

Figure 3.13 Percent increase in ln(Revisit Interval) when the patient sees a transactional provider($RD_i = 0$) compared to when the patient sees her regular doctor($RD_i = 1$), for patients at in urbal and rural areas

3.16.1 Effect of continuity of care and gender differences

Another source of health inequality is due to gender. Gender bias is prevalent in healthcare and can lead to damaging outcomes. According to Alspach (2012), gender bias is when patients might be assessed, diagnosed and treated differently for the same disease or complaint purely based on their sex. The study also assesses various literature and shows the wide spectrum across which gender bias exists, but ultimately leads to lower quality of healthcare being received by females.

Some examples include cardiovascular disease and pain management. Evidence suggests that cardiovascular disease is the leading cause of death in women, even though there is no physical reason why women and men should have different mortality rates for this disease (Mehta et al. 2016). An experimental study found that for males and females who presented the same symptoms of chronic heart disease (CHD), the primary care physicians was more likely to diagnose CHD in men and be more certain about their diagnosis, compared to women (Sayer and Britt 1997). On gender bias in pain management, Hoffmann and Tarzian (2001) finds that even though women experience more pain, they are less likely to be treated for pain than men. The authors attribute this difference to cultural stereotypes that have been prevalent through different generations and have led to gender biases.

Though not comparable across the two different genders, in the UK, endometriosis is one condition for which treatments have been so delayed because of the symptoms being 'discounted' as normal pain by medical professionals that the National Institute for Health and Care Excellence (Nice) has had to publish new guidelines asking medical professionals to pay attention to women's symptoms when they complain about crippling pelvic and period pain (Boseley 2017).

These findings suggest that primary care doctors may subconsciously provide differential service to women and men. Hence the mechanisms or the nature of the relationship with the doctor, specifically the incentive, information and trust aspects, may not be similar across men and women. It is possible that a transactional provider is more likely to exhibit gender bias as he is not familiar with the patient, whereas seeing the regular physician might help the women avoid the negative consequences of gender bias.

Looking at figure 3.14, we find that both men and women benefit from seeing their regular provider (green line), but we do not find evidence of a differential effect between seeing a transactional or regular provider (blue line). In other words, the productivity gain for women is not statistically different from the productivity gain for men, suggesting that even if gender bias is prevalent in primary care, fortunately, it does not necessarily manifest itself in the revisit interval of the patients.

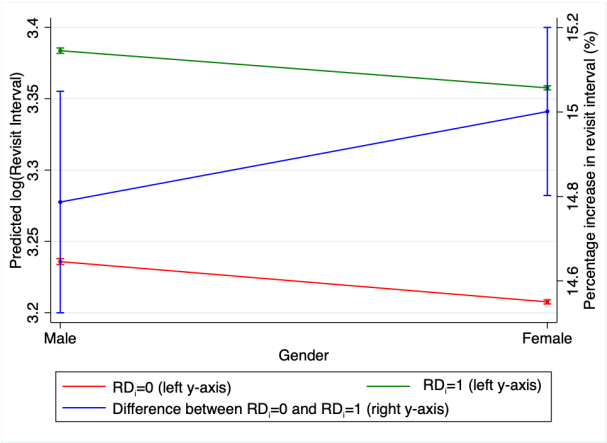

Figure 3.14 Both men and women benefit from seeing their regular provider, but the productivity gain for women is not statistically different from the productivity gain for men

Next, as gender and complexity are correlated, we explore whether males and females have statistically different productivity gains based on the number of comorbidities. Figure 3.15 combines the effect of providing care continuity to male and female patients with comorbidities ranging from 0 to 5+. The largest productivity enhancing effect comes from providing continuity to males with five or more comorbidities. For patients with 0-1 comorbidities, or for relatively less complex patients, the productivity gain from providing continuity to females is slightly higher than males, but for more complex patients with 2 or more comorbidities, the productivity gain for male and female patients is relatively similar regardless of the number of comorbidities.

We also investigate whether the moderating effect of deprivation index varies by gender. There

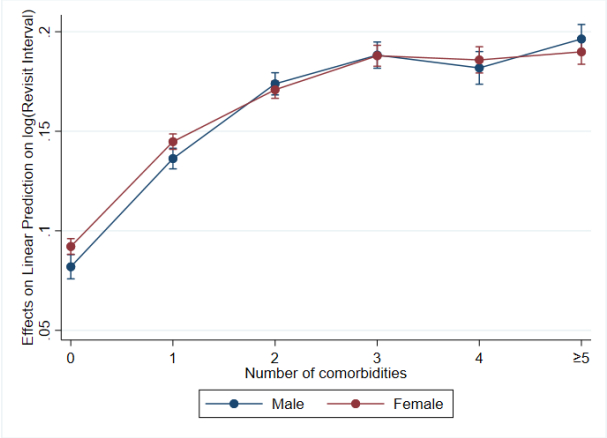

Figure 3.15 Differential effect of providing care continuity on RI for male and female patients across different comorbidities

has been documented evidence on systematic differences for both women and men across different socioeconomic groups (Allen and Sesti 2018). If women in the higher deprivation deciles benefit significantly more than the other groups, then offering continuity to this group of patients could be a potential way to narrow the gender and socio-economic inequality. But, looking at figure 3.16 we find that other than for the least deprived decile, for which the productivity gain for females is statistically larger than for males, the productivity gain between men and women for the other deprivation deciles is not statistically different.

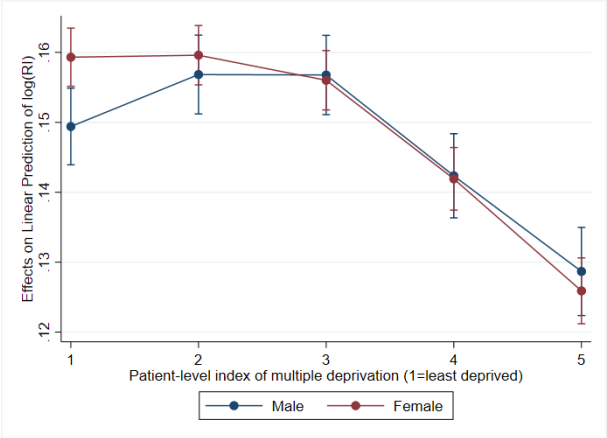

Figure 3.16 Other than for the least deprived decile, for which the productivity gain for females is statistically larger than for males, the productivity gain between men and women for the other deprivation deciles is not statistically different.

3.17 Percentage of past visits with regular doctor

To differentiate between patients who have had a larger share of their past consultations with their regular doctor and patients who have had a smaller share of their consultations, we include

$VisitPercentage$ as an additional control variable which is calculated at each consultation as the number of visits with the regular doctor in the 2 years preceding the focal consultation as a proportion of the total number of visits in the 2 years preceding the focal consultation. We find similar results after accounting for this potential confounder.

Dependent Variable = Natural logarithm of the revisit interval	
	Including $VisitPercentage$
RD_i=1	0.153***
	[0.151,0.154]
Observations	11,344,065

95% confidence intervals in square brackets

Table 3.17 Coefficient of RD_i for a model with $VisitPercentage$ control.

3.18 Doctor and patient heterogeneity

3.18.1 Patient-level heterogeneity

Patients might have their own unique circumstances that can affect their decision to revisit and when they are able to actually visit. To account for such patient level unobserved time-varying heterogeneity, we include patient random effects to our main model specification and find that the results remain similar.

Dependent Variable = Natural logarithm of the revisit interval	
	Patient RE
RD_i=1	0.131***
	[0.130,0.134]
Observations	11,344,065

95% confidence intervals in square brackets

Table 3.18 Coefficient of RD_i for a model with patient-level random effects.

3.18.2 Doctor specific effects

According to Schwartz et al. (1999), physicians receive very little guidance on visit scheduling times and hence widely vary in their recommendations of visit intervals. To capture physician specific variation and scheduling preferences, we include physician fixed effects (instead of practice fixed effects) to our main model. Additionally, we also believe that physician roles can be a confounder. For example, partners co-own the practice, hence have greater responsibility towards their patients, and have financial liability as owners of the practice. Salaried physicians, on the other hand, do not have a stake in the ownership and receive a fixed salary. Hence, partners may be more confident in scheduling a longer revisit interval. In both cases, we find similar results.

Dependent Variable = Natural logarithm of the revisit interval	
	Physician FE
RD_i=1	0.146***
	[0.144,0.148]
Observations	11,344,065

95% confidence intervals in square brackets

Table 3.19 Coefficient of RD_i for a model with physician-level fixed effects.

Dependent Variable = Natural logarithm of the revisit interval	
	Physician role
RD_i=1	0.139***
	[0.137,0.141]
Observations	11,344,065

95% confidence intervals in square brackets

Table 3.20 Coefficient of RD_i for a model with physician roles.

3.18.3 Practice−year fixed effects

To account for the changes in practice scheduling policy over time, we additionally include Practice−Year fixed effects and find that the results remain similar.

Dependent Variable = Natural logarithm of the revisit interval	
	Practice−Year fixed effects
RD_i=1	0.154***
	[0.152,0.156]
Observations	11,227,134

95% confidence intervals in square brackets

Table 3.21 Coefficient of RD_i for a model with practice year fixed effects.

References

Ahuja V, Alvarez CA, Staats BR (2020) Maintaining continuity in service: An empirical examination of primary care physicians. *Manufacturing & Service Operations Management* .

Allen J, Sesti F (2018) Health inequalities and women – addressing unmet needs. *British Medical Association* 1–12.

Alspach JG (2012) Is there gender bias in critical care? *Critical Care Nurse* 32(6):8–14, URL `http://dx.doi.org/10.4037/ccn2012727`.

Barker I, Steventon A, Deeny SR (2017) Association between continuity of care in general practice and hospital admissions for ambulatory care sensitive conditions: cross sectional study of routinely collected, person level data. *Bmj* 356.

Baum CF, Schaffer ME, Stillman S (2002) IVREG2: Stata module for extended instrumental variables/2SLS and GMM estimation. Statistical Software Components, Boston College Department of Economics, URL `https://ideas.repec.org/c/boc/bocode/s425401.html`.

Bobroske K, Freeman M, Huan L, Cattrell A, Scholtes S (2021) Curbing the Opioid Epidemic at its Root: The Effect of Provider Discordance after Opioid Initiation. *Forthcoming in Management Science* .

Boseley S (2017) 'Listen to women': UK doctors issued with first guidance on endometriosis.

Duan N (1983) Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association* 78(383):605–610.

Dugdale DC, Epstein R, Pantilat SZ (1999) Time and the patient–physician relationship. *Journal of General Internal Medicine* 14(Suppl 1):S34.

Fisher R, Dunn P, Gershlick B, Asaria M, Thorlby R (2020) Level or not? Technical report, URL `http://dx.doi.org/10.37829/HF-2020-RC13`.

Freeman M, Robinson S, Scholtes S (2020) Gatekeeping, fast and slow: An empirical study of referral errors in the emergency department. *Management Science* .

Gulliford MC, Sun X, Anjuman T, Yelland E, Murray-Thomas T (2020) Comparison of antibiotic prescribing records in two uk primary care electronic health record systems: cohort study using cprd gold and cprd aurum databases. *BMJ open* 10(6):e038767.

Hill AP, Freeman GK (2011) Promoting continuity of care in general practice. *London: Royal College of General Practitioners* .

Hjortdahl P, Borchgrevink CF (1991) Continuity of care: influence of general practitioners' knowledge about their patients on use of resources in consultations. *British Medical Journal* 303(6811):1181–1184.

Hoffmann DE, Tarzian AJ (2001) The girl who cried pain: a bias against women in the treatment of pain. *Journal of Law, Medicine & Ethics* 29(1):13–27.

Irving G, Neves AL, Dambha-Miller H, Oishi A, Tagashira H, Verho A, Holden J (2017) International variations in primary care physician consultation time: a systematic review of 67 countries. *BMJ Open* 7(10):e017902.

Jeffers H, Baker M (2016) Continuity of care: still important in modern-day general practice.

Kontopantelis E, Olier I, Planner C, Reeves D, Ashcroft DM, Gask L, Doran T, Reilly S (2015) Primary care consultation rates among people with and without severe mental illness: a uk cohort study using the clinical practice research datalink. *BMJ Open* 5(12).

Leniz J, Gulliford MC (2019) Continuity of care and delivery of diabetes and hypertensive care among regular users of primary care services in chile: a cross-sectional study. *BMJ Open* 9(10).

Maarsingh OR, Henry Y, van de Ven PM, Deeg DJ (2016) Continuity of care in primary care and association with survival in older people: a 17-year prospective cohort study. *British Journal of General Practice* 66(649):e531–e539.

MacCallum RC, Zhang S, Preacher KJ, Rucker DD (2002) On the practice of dichotomization of quantitative variables. *Psychological methods* 7(1):19.

Mehta LS, Beckie TM, DeVon HA, Grines CL, Krumholz HM, Johnson MN, Lindley KJ, Vaccarino V, Wang TY, Watson KE, Wenger NK (2016) Acute Myocardial Infarction in Women : A Scientific Statement from the American Heart Association. URL `http://dx.doi.org/10.1161/CIR.0000000000000351`.

Mercer SW, Fitzpatrick B, Gourlay G, Vojt G, McConnachie A, Watt GC (2007) More time for complex consultations in a high-deprivation practice is associated with increased patient enablement. *British Journal of General Practice* 57(545):960–966.

Nuffield Trust (2018) Poor areas left behind on standards of GP care, research reveals.

Olthof M, Groenhof F, Berger MY (2019) Continuity of care and referral rate: challenges for the future of health care. *Family Practice* 36(2):162–165.

Palmer W, Hemmings N, Rosen R, Keeble E, Williams S, Imison C (2018) Improving access and continuity in general practice. *Research Summary* .

Payne RA, Mendonca SC, Elliott MN, Saunders CL, Edwards DA, Marshall M, Roland M (2020) Development and validation of the cambridge multimorbidity score. *CMAJ* 192(5):E107–E114.

Pollack CE, Hussey PS, Rudin RS, Fox DS, Lai J, Schneider EC (2016) Measuring care continuity: a comparison of claims-based methods. *Medical care* 54(5):e30.

Rosen R, Massey Y, Abbas S, Hufflett T (2020) Relational continuity for general practice patients with new and changing symptoms. Technical report, Valentine Health Partnership, The Health Foundation.

Sanderson E, Windmeijer F (2016) A weak instrument f-test in linear iv models with multiple endogenous variables. *Journal of econometrics* 190(2):212–221.

Sayer GP, Britt H (1997) Sex differences in prescribed medications: Another case of discrimination in general practice. *Social Science and Medicine* 45(10):1581–1587, ISSN 02779536, URL `http://dx.doi.org/10.1016/S0277-9536(97)00095-6`.

Schwartz LM, Woloshin S, Wasson JH, Renfrew RA, Welch HG (1999) Setting the revisit interval in primary care. *Journal of General Internal Medicine* 14(4):230–235.

Stock JH, Yogo M, et al. (2005) Testing for weak instruments in linear iv regression. *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg* 80(4.2):1.

Tammes P, Purdy S, Salisbury C, MacKichan F, Lasserson D, Morris RW (2017) Continuity of primary care and emergency hospital admissions among older patients in england. *The Annals of Family Medicine* 15(6):515–522.

The Nuffield Trust (2020) Poorest get worse quality of NHS care in England, new research finds.

Todd A, Copeland A, Husband A, Kasim A, Bambra C (2014) The positive pharmacy care law: an area-level analysis of the relationship between community pharmacy distribution, urbanity and social deprivation in england. *BMJ open* 4(8):e005764.

Verma P, Ford JA, Stuart A, Howe A, Everington S, Steel N (2016) A systematic review of strategies to recruit and retain primary care doctors. *BMC Health Services Research* 16(1):1–25.

Williams E, Buck D, Babalola G (2020) What are health inequalities? *The King's Fund* (February 2020):n/a.

Worrall G, Knight J (2011) Continuity of care is good for elderly people with diabetes: retrospective cohort study of mortality and hospitalization. *Canadian Family Physician* 57(1):e16–e20.

Chapter 4

The Operational Determinants of Relational Continuity of Care

In this chapter we use the same detailed clinical data set from the Clinical Practice Research Datalink (CPRD) to investigate which factors affect a primary care practice's ability to provide RC to its patients. The study examines this question from an operations management perspective by exploring the relative importance of two operational factors that may explain variation in rates of RC between practices and over time: workload changes and workforce fragmentation. It is the first study to our knowledge that provides the first empirical study on the antecedents of RC in the primary care context. We find that a sustained increase in workload - caused by demand growth - and increasing fragmentation of the workforce - due to a shift to part-time and agency work - induces significant heterogeneity between practices in their ability to provide RC. In fact, these factors alone can also explain more than 50% of the decline in RC over the past decade, with workforce fragmentation having a relatively greater impact than demand growth. Using the insights from this analysis, practice managers can be aware of the root cause of low rates of RC at their own practice and the key operational levers that they can use to promote RC. Specifically, practice managers must improve the attractiveness of full-time salaried employment if they wish to promote RC. Meanwhile, strategies to curtail demand will be less effective and more challenging to achieve in practice.

4.1 Introduction

Primary care is facing a workforce crisis. Forecasts suggest that there will be a shortage of up to 55,000 primary care physicians (PCPs) in the United States (US) by 2033 (Heiser 2019). Meanwhile, the United Kingdom (UK) was already estimated to already have 2,500 fewer full-time equivalent PCPs than needed in 2019, with this shortage projected to almost triple by 2024 (Beech et al. 2019, Palmer 2019). The shortages in the labor supply are being compounded by chronic excessive workload stemming from an aging population with increasingly complex health needs(Kings Fund 2015). These trends pose a critical challenge for primary care managers: How can a stagnant or shrinking clinical workforce manage the increasing demand for primary care services?

Most primary care practices have sought to meet this challenge and satisfy the rising demand

for consultations by improving daily throughput and offering more consultations per clinician day (Boyle et al. 2010, Institute for Government 2019). While these actions have been partially effective in combating the growing demand for consultations, they also appear to have taken a toll on the workforce. Stress, absenteeism, and burnout are endemic in the industry, driving many clinicians to leave the profession or transition to part-time or more flexible working practices (Baird and Holmes 2019). Moreover, recent trends indicate that waiting times for primary care are on the rise and patient satisfaction is declining (Institute for Government 2019, Gault 2019). These developments suggest that current strategies for addressing systemic workload and workforce issues in primary care may, on their own, no longer be sufficient.

In light of these challenges, primary care practices might consider adopting another approach to mitigate the growing pressures: reducing demand by keeping their patients as healthy as possible and pro-actively reducing their need for consultations. However, in a system already strained, it is not immediately clear how to improve care provision and reduce demand without providing patients with *more* care, at least in the short term. A recent paper by Kajaria-Montag et al. (2020), however, suggests that a potential solution may be to not increase the resources dedicated to patient care but to instead focus on *who* is providing that care. In particular, the authors show that when a patient is seen by the doctor with whom they are most familiar, they have shorter consultation times and return later for a subsequent visit. Their findings suggest that practices that prioritize continuity of care (COC) may be significantly more effective at reducing demand for consultations and improving the productivity of their clinical workforce.

Importantly, COC not only has the potential to improve clinical productivity but also has been shown to drive better patient health outcomes (e.g., lower mortality rates) and improve system performance (e.g., reduce ED visits). These advantages are perhaps not surprising since, in contrast to the episode-focused secondary care model in which "diseases stay and patients come and go," in the primary care setting "patients stay and diseases come and go" (Heath 1995). Repeated interactions can thus increase a PCP's sense of ownership and personal responsibility for their patients' health and well-being, improving clinical decision making and job satisfaction (Grembowski et al. 2005). Additionally, with health systems and governments increasingly trialing and adopting integrated care models to reduce costs and improve efficiency, primary care practices that are better at providing COC may come to find that they are rewarded for delivering better patient outcomes and system performance. Given the various potential benefits of COC for patients, providers, and health systems, the central question of the present study is therefore: What levers are available to a primary care practice manager looking to deliver COC at their practice?

It turns out that the answer to this question is not so straightforward. Surprisingly, there is an absence of practitioner-oriented advice and a lack of evidence-based academic research to inform strategies to promote COC. A 2010 evidence review by The King's Fund, a UK-based health policy think-tank, for example, concluded: *"We were struck by the absence of agreed policies or any general body of expertise on how to encourage continuity. Specific guidance is*

also lacking from the [professional membership body for family doctors in the UK]. Meanwhile, many developments in practice and national policy have had the unintended consequence of making relationship continuity more difficult to achieve" (King's Fund, 2010, p. 51). And while practitioners, professional bodies, and policymakers have shown a renewed interest in COC of late, there is still an absence of evidence as to which factors affect a practice's ability to provide COC.[1] This paper sets out to fill this gap in the literature by identifying strategies that primary care practice managers can implement to promote care continuity within the constraint of a diminishing workforce and to increase their awareness of the root causes of low continuity.

Specifically, the present study focuses on critical operational levers that may affect a practice's ability to provide COC: the composition of the clinical workforce and workload. While workload pressure in healthcare has been thoroughly studied from the perspective of demand-side issues – e.g., population growth and an older patient pool with more complex morbidities – supply-side issues have recently emerged as an equally worrying trend. Moreover, in contrast to the workforce size, which is difficult to expand due to ongoing labor shortages, the composition of the workforce can be actively managed. For example, practice managers have the autonomy to customize their service offering and can design attractive salary structures to affect their workforce composition.

We hypothesize that an increase in the number of patients, an increase in the consultation rate per patient, an increase in the proportion of part-time work and agency work, all lead to a reduction in continuity of care. Our empirical study leverages a rich dataset from the Clinical Practice Research Datalink (CPRD), which is an extensive database of anonymized patient-level primary care electronic health records from a network of primary care practices across the UK (Herrett et al. 2015). This longitudinal dataset contains detailed information on $\simeq$970 million primary care consultations for 5,686,257 patients (approximately 10% of the population) and 407 primary care practices in England between 2007 and 2017. From this, we construct a monthly panel that captures the extent to which each practice can provide patients access to their regular primary care provider over time.

Using advanced panel estimation techniques – in particular, autoregressive distributed lag (ARDL) models, which for each factor in our study separate the long- and short-run impact on RC while addressing possible endogeneity bias – we investigate the impact of two workload-related and two workforce-related factors on RC. Concerning workload, we find that (i) an increase in the practice population by one between-practice standard deviation (BPSD) leads to a 7.4 percentage point (p.p.) reduction in rates of RC (a 17% relative reduction from the mean RC rate). We also find that (ii) a one BPSD increase in the rate of consultations per patient leads to a 3.1 p.p. decrease in RC levels (7% relative). Meanwhile, greater dependence on a fragmented workforce consisting of (iii) part-time doctors and (iv) locums, each increasing by one BPSD, results in

[1] Recently, scholars in the healthcare operations management (OM) community have also begun exploring the health and system benefits associated with COC. However, as we will discuss, this stream of research predominantly focuses on the consequences rather than the antecedents of COC.

a 3.5 p.p. fall in RC provision for both (8% relative). Thus, our results suggest that patients who value access to their regular doctor should look to register with smaller practices staffed primarily by full-time salaried (i.e., non-locum) doctors. In fact, we estimate that attending a practice in which these four factors take values one BPSD below the mean would allow patients to see their preferred provider ≃61% of the time, more than double the ≃26% rate when these factors take values one standard deviation above the mean.

These results can also help to explain the decline in RC provision over time. Consistent with measures self-reported by patients, our data show a significant reduction in patient-provider continuity, from ≃48% in January 2008 to ≃36% in December 2017, a relative decrease of ≃25%. In explaining this trend, we find, contrary to expectation, that the most important within-practice drivers of RC change are related to workforce fragmentation rather than workload increases. Specifically, we show that a shift to a more fragmented workforce that comprises more part-time salaried practitioners and is more dependent on locums can explain ≃33% of the fall in RC. By contrast, the sustained increase in workload can explain relatively less of the variation in RC: When we also control for the change in the number of patients registered at a practice and the rate of consultations per patient, the proportion of the decline explained increases from ≃33% to ≃48%. Our results therefore suggest that to provide higher levels of RC for their patients, practice managers should focus on countering trends by hiring and maintaining a core workforce of full-time workers rather than depending on part-timers and locums. Where workforce fragmentation is unavoidable, managers should instead find strategies to counteract its detrimental effects on RC.

We comment further on the implications of this work and possible strategies to mitigate the adverse effects of workload and workforce pressures on RC in Section 4.7.

4.2 Literature Review

The primary contribution of this paper is to the operations and healthcare management literature relating to COC. (Note that we use the term COC in the literature review section because it is used by most other studies, though our study focuses on the RC component of COC.) As we will see, although the extant empirical literature on COC is extensive, most studies focus on the consequences rather than antecedents of COC. In addition, our insights are relevant to the operations literature focusing on dedicated queuing disciplines and the advantages of repeated interactions with the same server. In this section, we outline how our paper's contributions are positioned within these literature streams.

4.2.1 Operations Literature

Customer-server continuity.

COC has been studied as an essential driver of different operational and health outcomes in the operations management literature. For instance, Ahuja et al. (2020b) examine the association between COC in primary care and the frequency of secondary care inpatient visits, inpatient LOS, and hospital readmission rates for chronic diabetes patients, finding an inverted U-shaped relationship. The effects are also found to be more pronounced for more complex patients. The authors follow up on this study in Ahuja et al. (2020a) by showing the adverse effects of reducing COC on medication adherence and glycemic variability, observing that this can partially explain the negative impact of COC on patient outcomes. Senot (2019), meanwhile, extends the notion of COC beyond the relationship between individual providers and their patients to also include continuity across the physical location in which care is delivered and the organization that provides the care. In a study of heart failure patients, the author finds that all three forms of continuity are important in reducing patient readmission rates.

While the studies above establish the link between COC and important outcomes, other work has started to explore how and when COC delivery can be improved. Queenan et al. (2019), for example, find that technology-enabled COC, coupled with a higher level of patient involvement in their own care, can reduce hospital admissions for patients with chronic obstructive pulmonary disease. Meanwhile, Bobroske et al. (2020) point out that while COC is generally encouraged in the post-acute phase of treatment, there may be advantages of more fragmented care in the initial treatment stages, e.g., a greater diversity of provider opinion and the relative absence of cognitive biases like anchoring. They show that for new opioid initiates, provider discordance (rather than continuity) can reduce the likelihood that a patient becomes a long-term opioid user.

Repeated interactions have also been discussed outside of the healthcare domain. Evidence from finance, for example, suggests that since it is costly to search for service providers, repeated interactions between investment banks (the provider) and investors (the customer) can lead to favorable pricing of convertible bonds (Henderson and Tookes 2012). In the context of contracting between a firm and a supplier when the parties are not yet at the stage of writing court-enforceable contracts, repeated interactions (or anticipated repeated interactions) encourage the adoption of relational contracts (informal agreements) that are built on trust and cooperation (Taylor and Plambeck 2007). These studies above have looked at relationsips between organizations rather than people. Canales and Greenberg (2016) examine relational contracts in the context of dyadic relationships between people rather than organizations, and find that other than the consistency of a loan officer in the context of a micro-finance loan setting, the relationship style consistency helps mitigate the negative impact of the broken ties between the loan officer and the customer.

Unlike this existing work on customer-server continuity, which mostly shows the benefits across a range of outcomes and contexts, our paper instead explores the question of how service providers can provide greater continuity to their customers. Specifically, we focus on identifying the primary drivers of differences in rates between and declines of RC across primary care practices.

Pooled versus dedicated queues.

The notion of allocating a patient to their COC provider is also akin to deciding whether to operate a dedicated versus pooled queueing system, with patients more easily allocated to their preferred providers using a dedicated queueing approach.

Although pooled queueing is often used in healthcare, research from other contexts has highlighted that pooled queueing is especially poorly suited for contexts with non-identical servers (Smith and Whitt 1981) or when there are different customer classes (Benjaafar 1995). Examples of the detrimental effects of pooled queueing include the depersonalization of service, lower customer satisfaction, and less opportunity for server specialization. These effects can have counterintuitive results: In the context of call centers, one study found that when customers were grouped to be served by a dedicated team of agents, both speed and quality improved (Jouini et al. 2008). Social loafing, which occurs when service providers exert less effort because task accountability lies with a group rather than an individual, is one mechanism that has been used to explain the inefficiency of pooled queues. For example, in grocery store checkouts, Wang and Zhou (2018) find that dedicated queues are faster than pooled queues due to the social loafing effect, with pooling having a negative indirect effect on service time.

Prior work in healthcare operations has also highlighted several benefits of a dedicated queuing approach when customer needs are heterogenous. In the emergency department (ED) setting, for example, Saghafian et al. (2012) use an analytical and simulation approach to show that segregating ED beds and care teams based on the likelihood that patients will be admitted or discharged to hospitals can improve ED performance. Meanwhile, an empirical study by Song et al. (2015) shows that dedicated queueing configurations can reduce patients' LOS in the ED by increasing physicians' feelings of ownership over patients and resources. Building on this existing literature, in this paper we study RC as another benefit of the dedicated queueing setup in a system that features repeated interactions between customers and providers. Specifically, all primary care practices in our dataset operate some combination of a pooled and a dedicated system, wherein some patients are allocated to the next available doctor while others are able to request or are scheduled with their provider of choice. This paper contributes to the operations literature by identifying the main factors that can cause practices to shift from a more dedicated to a more pooled queueing system characterized by lower rates of RC.

4.2.2 Medical Literature

Similar to the work in the operations community, most of the medical literature on COC focuses on establishing the advantages of a long-term patient-doctor relationship (cf. the review of the continuity literature by Haggerty et al. 2003). Benefits include, for example, patients whose medications are prescribed by their COC doctor being more adherent and compliant with the medication (Dossa et al. 2017) and less likely to fill risky prescriptions, e.g., among opioid users (Hallvik et al. 2018). Care continuity has also been associated with a better overall quality of life in cancer patients (Drury et al. 2020) and patients with hypertension (Ye et al. 2016) and with reductions in mortality risk across a range of conditions (Maarsingh et al. 2016, Cho et al. 2015).

In addition to the direct benefits to patients, various secondary care outcomes are also affected by primary care continuity. A meta-analysis conducted by Huntley et al. (2014) involving participants from all OECD countries concluded that repeated interactions with the same healthcare professional reduced unscheduled secondary care usage. For instance, studies have found reductions in ED presentations and unplanned admissions for patients with serious mental illness (Ride et al. 2019) and patients aged 65 and over (Tammes et al. 2017, Katz et al. 2015), lower rates of hospital admission for ambulatory care sensitive conditions (Barker et al. 2017), and fewer preventable hospitalizations (Nyweide et al. 2013). In general, at a system level, lower continuity is associated with increased healthcare utilization and higher levels of healthcare spending, especially among older patients (Amjad et al. 2016).

Few studies, meanwhile, have discussed the question of how to provide or increase the level of COC to patients or the question of what factors impede providers from being able to maintain a satisfactory level of COC. Those that do have typically focused on the demographic factors that predict continuity, such as deprivation scores, education levels, and mental health status. (Kristjansson et al. 2013, Levene et al. 2018). The work closest to this paper is by Kristjansson et al. (2013), who perform a cross-sectional study of 137 primary care practices in Ontario, Canada, and find that several practice-related factors – such as the number of staff and opening hours – also predict lower levels of COC. Unlike these works, we exploit the panel structure of our data to demonstrate the causal impact that workload and workforce-related factors play in creating variations in RC between providers and over time. Importantly, we find that it is these operational factors and not patient demographics that are most important in explaining RC variation.

4.3 Context and Hypothesis Development

This section provides background information on the primary care context in the UK, which is the focus of this study, before outlining the main hypotheses that this paper sets out to investigate.

4.3.1 Overview of Primary Care Provision in England

GP services.

In the UK, primary care is the standard point of entry to the health system, with GP practices (the UK term for a primary care practice) in England providing approximately 300 million consultations per year, more than ten times the number of visits to emergency departments (EDs) (NHS England 2018). Although the UK operates a publicly-funded healthcare system, GP practices are mostly owned and operated by doctors (i.e., GPs) and run as unlimited liability partnerships that act as independent contractors to the National Health Service (NHS). Thus, while most GP practices are similar in terms of their operations, their ownership structure means that they have a significant degree of autonomy. Most of the income of GP practices in England is determined by a standard contract with the English NHS under a capitation payment model, i.e., a fixed fee per registered patient per year. GP practices are generally not allowed to offer private clinical services to their registered patients.

A primary care practice has to accept every patient within a prescribed catchment area but can choose to accept or decline patients who apply from outside this area. A patient, meanwhile, can only register with one primary care practice. Coverage is nearly universal, with 98% of the UK population registered with a primary care practice (NHS England 2012). A patient who registers at a practice will be *administratively* assigned to a specific GP, referred to as the patient's "named GP." This is, however, a purely administrative requirement to reassure the patients that they have one GP who is responsible for their care, with patients entitled to see any GP employed by the practice at which they are registered. Patients can therefore choose a preferred GP (who often will not be their named GP) who, after repeated consultations, will begin to take responsibility for the health of that patient.

For patients, a visit to a GP is free at the point of care. Appointments are normally booked in advance – either in person, on the phone, or online – with the average wait time for a GP consultation (including both scheduled and urgent consultations) of approximately 13 days in 2016 (Gault 2019). Consultations themselves may be performed face-to-face, via telephone or, occasionally, at the patient's home. Assignment of patients to GPs for these consultations can depend on a variety of factors, including the preference of the patient, the availability of GPs, the degree of urgency, a patient's willingness to wait, and scheduling norms at the primary care practice.

For more urgent health concerns, for example an overnight asthma flare-up, patients can also access on-the-day GP services, which are delivered differently across practices. Some practices reserve a number of appointment slots for urgent services and, when those slots are fully booked, refer patients to the ED or book them for the next day. Some practices, meanwhile, only accept urgent patients who call in before a certain cutoff time, whereas other practices offer unlimited access for acute care throughout the day and have GPs dedicated to these urgent services.

In some instances, the GP may be unable to diagnose or treat the patient's needs within the primary care setting. They might then refer the patient for outpatient services or, if necessary, send them to the ED of a nearby hospital. (Note that the patient may also circumvent primary care entirely and go directly to the ED.) This gatekeeping function helps to preserve limited and expensive downstream resources for patients with the greatest need. Information from secondary care is fed back to the GPs, and the information is documented in their electronic health records system. Hence, primary care practices hold comprehensive and longitudinal medical records for their patients.

Primary care practice staffing and work patterns.

GP practices are commonly staffed by a combination of medical doctors (i.e., GPs), nurses, pharmacists, and other patient carers as well as administrative and clerical staff, who all play an important role in providing effective service to their patients and community. A typical primary care practice with 10,000 registered patients may have five full-time equivalent (FTE) GPs, five FTE nurses and other patient carers, and ten FTE administrators or clerical staff, who are overseen by a practice manager. However, size and workforce composition varies significantly between practices (Centre for Workforce Intelligence 2014).

During one full work day, a GP will usually perform at least 20 consultations of approximately 10-15 minutes each (Graham Clews 2013). While part-time GPs may perform fewer consultations in a day by working shorter hours, the norm in the profession is for part-timers to instead work fewer days per week. Some doctors may also perform certain tasks on their days off despite not being in the office, e.g., they may work on patient notes from home or follow up with patients by phone. When constructing the workforce-related variables later in Section 4.4.2.3, we must therefore be careful in defining what constitutes a full work day.

It is also important for our study that GPs working at these practices can be separated into two types who differ in their roles and responsibilities (NHS Improvement 2011):

- Established GPs: These are GPs who are under contract with a specific GP practice (either as partners/owners or as salaried practitioners).
- Unestablished GPs: These are professionals who are trained as GPs but who are not permanent employees at a particular GP practice. Instead, they are paid on a shift-by-shift basis for the work that they perform. They may, for example, be registered under a locum agency or else hold contracts with a number of GP practices simultaneously.[2]

Note that only established GPs at a practice are allowed to be listed as a patient's named GP. However, this does not prevent patients from having an unestablished GP as their preferred GP.

[2]Specifically, established GPs are defined to include senior partners, partners, salaried practitioners, and sole practitioners. Meanwhile, unestablished GPs are defined to include locums, GP registrars, and GP retainers.

4.3.2 Workload-Related Factors Affecting Relational Continuity

Several factors have contributed to increasing workload pressures faced by primary care practices in the UK. First, population growth by 6.4% between 2010 and 2018 (Office for National Statistics 2019) has spurred an increase in demand for GP consultations by ≃9% over the same period (Institute for Government 2019). This is occurring during a time at which there has also been a major restructuring and consolidation of primary care in the UK, with more than 1,000 practices covering over 4.2 million patients closed or merged between 2013 and 2018 (Bostock 2018), contributing to an overall reduction in the number of practices by 18% between 2004 and 2019. Consequently, the remaining GP practices have had to provide care to more patients, with the average size of a practice's patient list rising by ≃45% from ≃5,900 patients in 2004 to ≃8,500 in 2019 (Bostock 2019b).

Second, with most developed countries experiencing increasing life expectancies, the number and proportion of older patients are growing rapidly. In the UK, recent projections state that the number of people aged over 75 will increase from one in eleven in 2019 to closer to one in seven by 2040 (Tammes et al. 2019). At the same time, these patients are increasingly living with chronic diseases and multimorbidities and are placing an ever-increasing demand on primary care services, with patients with chronic illnesses estimated to account for approximately half of all GP consultations (Kings Fund 2015). As a result, in addition to an increase in the number of patients registered at each practice, the crude annual consultation rate per person grew by an estimated 10.5% from 4.7 in 2007–08 to 5.2 in 2013–14 (Hobbs et al. 2016).

Third, demand growth has also contributed to supply-side issues that have further exacerbated the workload pressures. In particular, many GPs are leaving the profession or reducing their work hours, while recruiting new trainees is increasingly challenging. This has been explained by low job satisfaction caused by less time spent with patients and a shift in focus away from patient-centered care (Doran et al. 2016). This is particularly well-documented in the US context, with 55% of physicians now describing their morale as negative (The Physicians Foundation 2018). Additionally, fatigue and burnout are becoming a major issue, with self-reported measures of burnout among US physicians increasing from 45% to 55% in just three years between 2011 and 2014. Meanwhile, low salaries and disputes over retirement benefits have only made matters worse (Baird and Holmes 2019). To avoid such unappealing working conditions, medical students are gravitating more towards specialist training and away from primary care or generalist training (Dalen et al. 2017). As a consequence, the size of the established GP workforce is stagnant or decreasing and not keeping up with the growth in demand (Palmer 2019).

In terms of the impact of these three trends on GP practice operations, note that the classic speed-quality trade-off in queuing systems tells us that in the face of growing demand, without a commensurate increase in supply, it is not possible to maintain both speed of access and quality of service (Anand et al. 2011). Primary care practice managers thus face an important choice in providing care and managing patient expectations as they search for a new trade-off point on the

new speed-quality curve. Governments, meanwhile, have been urging and incentivising primary care providers to improve speed of access, especially for patients with urgent needs, in order to relieve pressure on hospital emergency departments (Boyle et al. 2010). This prioritization of speed appears to have been taken seriously by practice managers, with patient surveys indicating an increase in the percentage of same-day appointments with a GP or nurse between 2012 and 2017 (Institute for Government 2019).

To provide speedy access and maintain clinical quality with a GP workforce that is not growing to keep up with demand thus requires more flexible scheduling practices. In line with basic queueing theory, one lever that can be pulled to reduce or maintain waiting times in the face of an increase in workload is increased pooling of some activities (Cachon and Terwiesch 2011). In particular, in a multi-server system, moving from a more dedicated queuing discipline (in which patients join the queue for a particular GP) to a more pooled setup (in which patients join a common queue and are allocated to the next available GP) can improve speed of access, all else being equal. This works by reducing the so-called "idle server" problem, which occurs when a queue builds for one GP but another (perhaps less popular) GP is available but has no work to perform.[3]

However, as discussed in Section 4.2.1, while pooling may be effective in reducing waiting times, in knowledge-intensive services such as primary care, pooled queue configurations make it harder for patients to access their regular providers. Consistent with this notion, over the same time frame as the aforementioned workload increases have occurred, the proportion of patients reporting as being able to see their preferred GP at least "most of the time" has decreased drastically, from 77% to 50% between 2009 and 2018 (Institute for Government 2019). While it has been proposed that workload factors are partially to blame for this fall in RC, e.g., due to the need for greater pooling to manage this higher workload, there is no empirical evidence linking these two phenomena. We therefore test the hypothesis that the level of RC will be lower at practices and at times where there are more registered patients and a higher consultation rate per patient.

Hypothesis 1. (H1) Given a fixed number of established GPs, an increase in the number of registered patients will reduce the level of relational continuity

Hypothesis 2. (H2) Given a fixed number of established GPs, an increase in the consultation rate per patient will reduce the level of relational continuity.

4.3.3 Workforce-Related Factors Affecting Relational Continuity

While the overall increase in primary care workload is well-known, intensifying workload pressures have also had an impact on the composition of the workforce. As noted above, low job

[3]The reality is, of course, that GPs are not idle but instead perform non-patient-facing duties (e.g., administrative tasks) or work slower (i.e., they use their discretion over service time). Note that the latter will not necessarily improve quality because the reason for spending more time with a patient is not driven by patient need.

satisfaction and burnout are causing some primary care providers to leave the profession. Others, meanwhile, have responded by shifting to part-time work or choosing portfolio careers, in which clinicians take on other roles such as management tasks or running pain clinics in addition to clinical work (Baird and Holmes 2019). Overall, the trend towards part-time work has been steadily increasing, and fewer than 30% of GPs in the UK now report working full-time (Bostock 2019a). With GPs increasingly preferring the flexibility and reduced responsibility that comes with working on an ad-hoc basis, evidence suggests that the number of GPs leaving their established positions and working instead as locums (i.e., unestablished GPs) has also been growing steadily over time (General Medical Council 2018). Thus, although historically established GPs have performed the bulk of the work, many practices now report relying on the unestablished workforce to fill at least a quarter of shifts (Matthews-King 2015).

While flexible work hours and growth in the unestablished workforce have helped to counteract declines in staffing numbers, they have also resulted in a more fragmented workforce. In particular, part-timers will not be present on all days of the week, while locums rotate between GP practices to fill shortages, meaning that they work few days per month at any one practice. For the approximately four in ten GP appointments that take place on the same day on which they are scheduled (Legraien 2019), intuitively, the likelihood that a patient's preferred GP will be working that day is lower if that GP works part-time or is a locum. Thus, significant variation in RC rates across practices can arise due to heterogeneity in their ability to retain a core workforce and formally prioritize continuity in a coordinated manner. This is not just hypothetical: Practices themselves have reported that use of part-time workers and locums has served to undermine service continuity and stable working conditions (NHS England 2016b).

Overall, this suggests that a practice will be less able to provide RC to their patients when relying more on (i) part-time workers and (ii) the unestablished workforce. We test this hypothesis by examining the degree to which these two factors affect RC provision across practices and time.

Hypothesis 3. (H3) An increase in the proportion of part-time work within the established workforce will reduce the level of relational continuity.

Hypothesis 4. (H4) An increase in the unestablished workforce as a proportion of the overall GP workforce will reduce the level of relational continuity.

4.4 Data and Variable Descriptions

This section provides detailed descriptions of the dataset and main variables for this study.

4.4.1 Data Preparation

4.4.1.1 Data description.

For this retrospective analysis of GP consultations, we collect data from the Clinical Practice Research Datalink (CPRD), which is a large database of anonymized patient-level primary care electronic health records from a network of GP practices across the UK. CPRD's database encompasses longitudinal data for over 11.3 million patients from 674 practices, found to be representative of the UK population in terms of age, sex and ethnicity (Herrett et al. 2015). The CPRD database contains a wealth of data on patients, providers, diagnoses, treatments, referrals, and more. Most of this information is provided in the form of codes, which can be used to categorize information on each patient visit and to construct our analysis sample.

From this database, data for the study includes all information for patients who had at least one primary care consultation between January 1, 2008 and December 31, 2017. This was narrowed by our data provider to (i) patients for whom it was possible to gain additional linked data from other health providers (such as secondary care) and (ii) practices in England that had consented to linkage. We thus obtained a comprehensive dataset of $\simeq$970 million primary care consultations for 5,686,257 patients and 407 practices. We also note that CPRD only includes practices that meet data quality standards, and so consultations were excluded if they occurred before the practice deemed the data to be of research quality.

In forming our sample, we further restrict it to consultations performed by a GP (rather than, e.g., nurse-led consultations) since our measure of RC is calculated at the GP-level (i.e., it is based on whether or not the patient had an appointment with their regular GP). This restriction is consistent with other literature that examines continuity in primary care settings (e.g., Tammes et al. 2017, Barker et al. 2017). It also results in a natural subset of consultations to analyze because the advantages of RC are less clear for other types of appointments, e.g., blood tests and vaccinations administered by nurses. Selecting only GP consultations reduces our sample to $\simeq$370 million observations.

Next, after discussion with a number of GPs and in accordance with other literature (e.g., Salisbury et al. 2009), we further restrict our sample to only include face-to-face visits. These represent 52% of all GP consultations in our sample and are the standard mode of patient-provider interaction for a new complaint or for ongoing care management. In particular, we discard telephone consultations, which are typically used for sharing test results or to triage patients; home visits, which are more common for vulnerable and seriously ill patients; and non-clinical (e.g., administrative) consultations, which are not the primary focus of our study. This leaves $\simeq$190 million consultations that we take forward for analysis.

4.4.1.2 Unit of analysis.

Recall our research objective is to establish the effect of workload and workforce factors on a GP practice's ability to provide RC to its patients. To perform this analysis while accounting for heterogeneity between practices, we therefore adopt a panel data structure by converting the consultation-level data to a monthly panel for each practice.

Note that some practices in the CPRD dataset transfer in or drop out of the dataset over our study window. Therefore, in forming the panel we exclude any primary care practice that was present in the sample for fewer than five years (i.e., half of the sample period).[4] This ensures sufficient monthly observations to estimate the effects within each practice reliably. This leaves a set of 320 practices (i.e., 79% of the 407 total) to be included in our sample, with each practice present for 96 months on average, yielding a total of 30,291 practice-month observations to be included in the analysis.

4.4.2 Variable Descriptions

We next describe the calculation of the key variables included in our study. All summary statistics are provided in Table 4.1.

4.4.2.1 Dependent variable.

To capture the extent to which a practice is able to deliver RC, we calculate the percentage of consultations that occurred between a patient and their regular GP in a given practice-month.

First, it is important to recall that a patient's named GP and their regular or preferred GP may not necessarily be the same. Previous studies have taken the view that it is the provider who the patient sees most regularly, and not the one to whom they are assigned, with whom they have greater familiarity, and hence the patient is more likely to benefit from repeated interactions with this regular provider (Senot 2019, Barker et al. 2017). Therefore, in defining our RC measure, we follow convention and consider whether a particular consultation was between a patient and their "regular GP" rather than their named GP.

Second, given the ten-year time horizon of our study, instead of treating a patient's regular GP as a fixed entity we will allow this to vary over time. This is important as various factors can lead to a change in a patient's regular GP, such as a GP leaving the practice or retiring, or a

[4]Since the panel is unbalanced, one might be concerned about non-random attrition from the sample, which could bias results. To check for this, we perform the sample selection test proposed by Verbeek and Nijman (1992). Results indicate that we do not reject the null hypothesis that attrition from the sample is random. We also repeat the analysis with only those practices that are continuously present in the CPRD dataset during the study period, giving us a cohort of 79 practices over ten years, yielding 9,480 practice-month observations. Results are consistent when estimated using this subsample and can be found in Section 5.3 of Chapter 5.

positive encounter between a patient and another GP that leads to a switch. Note that there is a risk in making this factor dynamic because a patient may change GPs so frequently that we cannot identify the regular GP with any certainty. We address this in our definition of the regular GP.

To identify the regular GP, observe that each consultation j will be associated with a particular patient i and occur at some time t, which we can denote $(it)[j]$. We define a patient's regular GP at consultation j as the GP that patient i saw more frequently across all face-to-face consultations with GPs over a two-year time window prior to time t.[5] The two-year time window ensures that the regular GP remains relatively stable from one consultation to the next.

Following convention in the medical and operations literature, if patient i had fewer than three consultations over the two-year window prior to time t then we exclude consultation j from the calculation of the dependent variable, since accurate identification of the regular GP is not possible (Ahuja et al. 2020b). This leaves $\simeq$117 million consultations, or an average of $\simeq$3800 per practice-month, to estimate our measure of monthly RC provision at each practice in our sample.

Next, we define a binary variable, $RegGP_{(it)[j]}$, which equals one if patient i sees their regular GP during consultation j and zero otherwise. We aggregate this measure to the practice-month level by averaging $RegGP_{(it)[j]}$ over the set C_{pm} of all GP face-to-face consultations that occur at practice p in month m and that offer sufficient information to identify the regular GP, i.e.:

$$RC_{pm} = \frac{\sum_{j \in C_{pm}} RegGP_{(it)j}}{|C_{pm}|}\,, \tag{4.1}$$

where $|C_{pm}|$ denotes the cardinality of set C_{pm}.

Figure 4.1 gives a histogram of the monthly RC measure as well as the trend over time. We observe a strong decline in RC over the time horizon of our study, as RC drops from an average of approximately 48% in early 2008 to nearer to 35% by the end of 2017.

4.4.2.2 Workload variables.

Our first two independent variables capture variation in levels of workload placed on a GP practice by measuring the number of registered patients and consultation rate per patient in a given month.

1. Number of registered patients. For each patient in our dataset, we have the date that they registered with a practice and, if applicable, the date on which they transferred out. Using

[5] Overall, 11% of cases result in a tie. In the case of a tie between an established and an unestablished GP, we pick the established GP because she is, by definition, more accountable for that patient as a salaried employee of the practice. For the remaining 7% of ties: i) if there are multiple established GPs amongst which a tie exists, then we randomly assign one as the patient's regular GP; ii) if there is a tie but none of the GPs are established GPs, then we randomly assign one of the unestablished GPs as the regular GP.

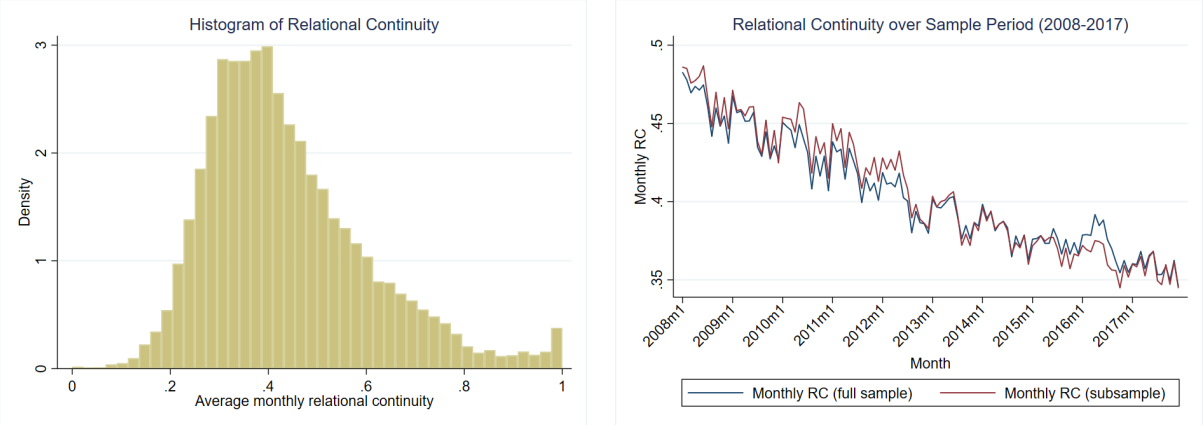

Figure 4.1 Histogram of relational continuity (left) and trend over time (right), with trend calculated using the weighted average of the full sample (blue line) and the subsample of practices operating during all ten years (red line).

these dates, we recreate the number of patients, $PracticePop_{pd}$, registered at each practice p on each day d. Averaging $PracticePop_{pd}$ for each practice p over all days d in month m gives us our measure of the monthly registered practice population, $PracticePop_{pm}$.

2. Use of GP services by registered patients. We take the count of the total number of face-to-face GP consultations at practice p in month m, $\left|C^{+}_{pm}\right|$, and divide through by the size of the patient list, $PracticePop_{pm}$, to measure use of GP services per patient in a practice-month, which we denote $ConsPerPat_{pm}$. Holding the practice population constant, an increase in this measure captures an increase in the frequency of primary care use by registered patients.

4.4.2.3 Workforce variables.

Two additional independent variables capture differences in workforce composition within and between practices.

3. Part-time work by established GPs. To measure this, let $EstGPDays_{pm}$ be the total number of full work days worked by established GPs at practice p in month m (see Section 4.3.1 for the definition of an established GP). Dividing this through by the total number of established GPs who worked at least one full work day in that practice-month, $EstGPs_{pm}$, gives us our variable of interest, $DaysPerEstGP_{pm}$. A decrease in this measure indicates that established GPs shifted to more part-time work or reduced working hours.

4. Dependence on unestablished GPs. We calculate the ratio between the number of full days worked by established GPs, $EstGPDays_{pm}$, and the number of full days worked by any GP (i.e., both established and unestablished GPs), $TotGPDays_{pm}$, at practice p in month m. This gives us a measure of the relative dependence on established GPs,

$ShareEstGPs_{pm}$. A decrease in this measure indicates a shift in activity away from the established workforce.

Note that constructing the workforce-related variables requires a definition of the total number of full days worked by GPs in a month. As highlighted earlier in Section 4.3.1, however, part-time work can take two forms: (i) working shorter hours within a day, and (ii) working fewer days during a week – the latter of which is more common in practice. Furthermore, even during non-working days, GPs might still, rarely, be recorded by the electronic health record system as having performed one or more consultations (e.g., due to coding errors). We therefore define a full work day as one in which a GP performed at least 10 consultations. This threshold is set high enough to eliminate coding errors and special cases (e.g., where a GP provides ad-hoc cover for one to two hours), while also being set sufficiently low to ensure that we can identify a full work day when one occurs (during which at least 20 consultations are typically performed). Using a 5 or 15 consultation cutoff instead does not change the results (see Section 5.4 of Chapter 5).

4.4.2.4 Control variables.

Various other factors might affect a practice's ability to provide RC and may confound the relationship between RC and the workload and workforce composition variables. The inclusion of practice fixed effects (FEs) in the model accounts for time-invariant factors that are specific to the practice: for example, whether it serves a rural or urban population, population socioeconomic status, etc. Time FEs, meanwhile, can adjust for any factors that change over time and have a common effect on all practices. However, FEs are unable to account for time-varying factors that differ across practices over time. Therefore, we have also defined a number of additional control variables at the practice-month level.

First, the workload-related hypotheses focus on the effect of changes in demand while requiring the supply of labor to be accounted for or fixed. We can proxy the available labor using the total number of days worked by GPs in a month, i.e., $TotGPDays_{pm}$. Since this will be highly correlated with demand (as a practice that offers more consultations will require more GPs working more days), to reduce multicollinearity concerns we divide $TotGPDays_{pm}$ through by $\left|C^{+}_{pm}\right|$ to give us a measure of the supply of labor relative to total demand in a practice-month, $GPDaysPerCons_{pm}$. Note that since the demand-side is already accounted for via the independent variables specified in Section 4.4.2.2, $GPDaysPerCons_{pm}$ captures the effect on RC provision of having more GP days available to provide consultations.[6]

Second, we also control for changes in service times by taking the average appointment duration across all C^{+}_{pm} consultations in a practice-month. This is an important control, as practices that

[6] Alternatively, we could control by taking the number of established GPs, $EstGPs_{pm}$, as a measure of supply, or by dividing through by the number of registered patients $PracticePop$ rather than by $\left|C^{+}_{pm}\right|$. Results are the same in sign and significance and very similar in size regardless of which approach we take.

are more successful at reducing service times will be able to treat more patients per day, which may help to counteract some of the anticipated negative effects of demand growth on RC.

Third, any change in patient population demographics within a practice over time may affect the practice's ability to provide RC. Therefore, we also control for the average patient age, average percentage of females, and average number of comorbidities per consultation, calculated over the set of all face-to-face GP consultations that took place in a practice-month, i.e., C^{+}_{pm}. The number of comorbidities assigned to a patient at each consultation is calculated using the Cambridge Multimorbidity Score (Payne et al. 2020), which is designed specifically for use with the CPRD database and uses the patient's past consultation history to identify the presence of 37 different conditions, such as hypertension, depression, diabetes, heart disease, cancer, and more.

Lastly, we control for the month of the year (e.g., January, February, etc.) to account for seasonality, since workload and workforce composition can differ throughout the year, such as during summer or winter holidays, flu season, etc., and might affect the practice's ability to provide RC.

4.4.3 Summary Statistics

Panel A of Table 4.1 contains summary statistics for each of the main variables described in Sections 4.4.2.1 through 4.4.2.3. A quality check of the data verifies that it is consistent with expectations. The average list size per practice, 8,252, is close to the 7,860 reported by NHS Digital (2017). Furthermore, a recent survey found that the average GP now works fewer than 3.5 days per week, or less than 15 days a month, which is close to the 12.8 days per month reflected in our data (Donnelly 2018).

We note that in Panel A of Table 4.1, the scale of the variables differs significantly. This can lead to matrix inversion issues when performing maximum likelihood estimation and can also make effect size comparisons challenging. To avoid such issues and ease interpretation, we standardize the independent variables and controls by taking their z-scores (i.e., subtracting the mean and dividing by the standard deviation). This is a linear transformation and so has no impact on the results, but coefficients in our models must now be interpreted as the impact on RC of a one standard deviation change in the corresponding variable.

Summary statistics for the standardized variables are reported in Panel B of Table 4.1, followed by a table of correlations in Panel C. The correlation table shows that (except for $zConsPerPat$), there is a moderate to strong degree of correlation between the independent variables and RC. This is especially the case for $zDaysPerEstGP$, for which the correlation with RC takes value 0.50 ($p < 0.001$). Meanwhile, the degree of correlation between the independent variables provides no cause for concern, with the variance inflation factors (VIFs) all taking values less than 1.24.

Table 4.1 Descriptive Statistics and Correlations for Variables

Panel A: Descriptive Statistics

					St. Dev.		
	Mean	Median	Min	Max	Overall	Between	Within
RC	0.44	0.41	0.00	1.00	0.17	0.14	0.08
$PracticePop$[a]	8.25	7.80	1.05	31.72	3.97	3.98	0.39
$ConsPerPat$	0.29	0.27	0.00	1.11	0.11	0.09	0.06
$DaysPerEstGP$	12.83	12.71	1.00	29.00	3.23	2.34	2.22
$ShareEstGP$	0.79	0.82	0.00	1.00	0.19	0.16	0.12

Panel B: Descriptive Statistics - Standardized Variables

					St. Dev.		
	Mean	Median	Min	Max	Overall	Between	Within
$zPracticePop$	0.00	-0.11	-1.81	5.91	1.00	1.00	0.10
$zConsPerPat$	0.00	-0.17	-2.69	7.63	1.00	0.84	0.57
$zDaysPerEstGP$	0.00	-0.02	-3.66	5.01	1.00	0.71	0.69
$zShareEstGP$	0.00	0.15	-4.06	1.10	1.00	0.81	0.61

Panel C: Correlations

	(1)	(2)	(3)	(4)	(5)
(1) zRC	1.00				
(2) $zPracticePop$	-0.34***	1.00			
(3) $zConsPerPat$	-0.01^{+}	-0.06***	1.00		
(4) $zDaysPerEstGP$	0.50***	-0.12***	0.29***	1.00	
(5) $zShareEstGP$	0.38***	-0.06***	-0.03***	0.30***	1.00

[a] $PracticePop$ reported in thousands; $^{+}p < 0.10$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$; St. Dev. short for standard deviation.

4.5 Fixed Effects Models

4.5.1 Fixed Effects Estimator

We organize the data into an unbalanced panel with two levels, practice and time (i.e., month) and estimate a time and entity FE regression model specified by the equation:

$$\begin{aligned} RC_{pm} &= \alpha_p + \gamma_m + \beta_1 PracticePop_{pm} + \beta_2 ConsPerPat_{pm} \\ &+ \beta_3 DaysPerEstGP_{pm} + \beta_4 ShareEstGPs_{pm} + \boldsymbol{\beta}_5^{\mathsf{T}} \mathbf{X}_{pm} + \epsilon_{pm} \,. \end{aligned} \tag{4.2}$$

Practice-specific intercepts, α_p, capture unobserved time-invariant heterogeneity across practices. Common time effects, γ_m, capture shocks and trends in RC which affect all practices in the sample. Meanwhile, the vector $\mathbf{X}_{pm}$ contains the set of control variables described in Section 4.4.2.4, and $\epsilon_{pm} \sim \mathcal{N}(0, \sigma^2)$ is the idiosyncratic error term. Standard errors are clustered at the practice level to account for autocorrelation within the same practice.

4.5.2 Results

Results from the FE regression are reported in Table 4.2. The top panel in the results table reports the coefficients of the variables of interest and the continuous controls. The bottom panel reports the structure of the controls that are included as FEs, with "Yes" or "No" indicating inclusion or non-inclusion, respectively. Columns (1)–(4) provide results where the main independent variables of interest are included one at a time, with time and entity FE controls only. In column (5), all four independent variables of interest are included simultaneously, again with time and entity FE controls only. Finally, column (6) gives the results from the full model in which all controls (e.g., age, gender mix) are added to the model in column (5). Additionally, we report the R^2, which is the residual variance in RC explained after accounting for practice FEs (i.e., the within-practice R^2). For comparison, the baseline model with only time FEs has residual R^2 equal to 0.161.

Since the addition of controls does not change the results significantly, we proceed to discuss the estimates from the fully specified model in column (6). Starting from the workload factors, all else remaining equal, a 1σ increase in $PracticePop$ leads to a 0.043 ($p < 5\%$) percentage point (p.p.) decrease in RC. In addition, all else being equal, a 1σ increase in $ConsPerPat$ leads to a 0.032 ($p < 0.1\%$) percentage point (p.p.) decrease in RC. With respect to the workforce factors, all else being equal, a 1σ decrease in $DaysPerEstGP$ leads to a 0.032 p.p. ($p < 0.1\%$) reduction in RC, in line with our expectation. In addition, a 1σ decrease in $ShareEstGP$ corresponds to a 0.034 p.p. ($p < 0.1\%$) decrease in RC.

Table 4.2 Fixed effects panel regression results.

	(1)	(2)	(3)	(4)	(5)	(6)
zPracticePop	-0.038^{+}				-0.047*	-0.043*
	(0.023)				(0.020)	(0.019)
zConsPerPat		-0.014***			-0.020***	-0.032***
		(0.004)			(0.004)	(0.005)
zDaysPerEstGP			0.033***		0.029***	0.032***
			(0.003)		(0.003)	(0.003)
zShareEstGP				0.048***	0.038***	0.034***
				(0.004)	(0.004)	(0.003)
zGPDaysPerCons						-0.019***
						(0.006)
zMale						0.004
						(0.003)
zAge						0.020*
						(0.008)
zComorbidity						-0.001
						(0.007)
zConsDuration						-0.006
						(0.004)
Practice FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Month of Year	No	No	No	No	No	Yes
Observations	30283	30283	30283	30283	30283	30283
R^2	0.163	0.169	0.223	0.281	0.324	0.348

Notes: $^{+}\,p < 0.10$, $^{*}\,p < 0.05$, $^{**}\,p < 0.01$, $^{***}\,p < 0.001$; Standard errors in parentheses, clustered by practice; R^2 specifies the within-practice variance in RC explained after accounting for between-practice variation with practice FEs, with practice FEs alone explaining 75% of the variation.

Due to the limitations of the FE approach, which we discuss next in Section 4.5.3, we postpone interpretation of the results for later. However, we note that by simply comparing the lift in the residual variance explained (i.e., the R^2) when each of regressors are added one at a time in columns (1)–(4) of Table 4.2, it appears that the workforce factors explain relatively more of the variation in RC than workload factors – we return to this observation later in Section 4.6.2.

4.5.3 Limitations

Although the FE estimator described above provides preliminary evidence that the variables of interest have an impact on RC, it also has a number of limitations that drive us to adopt an alternative modeling approach in Section 4.6. These include (1) the inability of the FE estimator

to account for non-stationarity, a concern in large macro panels; (2) dynamic misspecification of the model, especially in the presence of serial correlation; and (3) violation of the slope homogeneity condition, which occurs when time-varying factors affect practices differently. All of these issues are discussed in more detail in Section 5.2 of Chapter 5.

Dynamic effects are said to exist if past values of COC influence future values of the regressors (i.e., if COC_{py-1} affects x_{py}) or if current realizations of the dependent variable are influenced not only by the current value of the independent variables, but also by values of the dependent and independent variable(s) in the past (i.e., if COC_{py-1} affects COC_{py} or if x_{py-1} affects COC_{py}). In our setting, for example, it is conceivable that doctors become disengaged when there is a shift from more relational to more transactional work practices (i.e., as COC reduces), resulting in doctors burning out, reducing their work hours, or quitting the practice entirely. A fall in COC in the past may thus affect future values of the independent variables. If these dynamics are present but not sufficiently captured, coefficient estimates may be biased (Campos and Kinoshita 2008).

In addition to dynamic effects, endogeneity can bias the results in two other ways. The first occurs when there exist time-varying omitted variables correlated with both the dependent variable and the independent variables. For example, a practice that prioritises speed of access (a potential time-varying omitted variable) may have lower rates of COC, e.g., because of a greater tendency to exploit the queuing benefits associated with pooling. However, to provide timely access to care, this practice may also require more doctors and potentially a more varied staff-mix, thus affecting the workforce composition. The second source of endogeneity is simultaneity bias, which arises when the dependent and independent variables are co-determined. For instance, suppose a practice chooses their workforce with a view towards achieving a particular level of COC in the same period. Then while COC may be affected by the workforce composition, the workforce composition will also be affected by COC.[7]

The standard strategy for addressing omitted variable and simultaneity bias is to identify external (i.e., strictly exogenous) instrumental variables correlated with the regressors but orthogonal to the dependent variable. Unfortunately this is not a viable approach in the present study, since it would require identifying and justifying at least five external instruments (one for the lagged dependent variable and one for each of the four main independent variables). In most empirical studies it can be challenging to find an instrument for a single potentially endogenous regressor; thus, identifying five external instruments appears to be a dead end.

In addition, the FE estimator only estimates the short-run (SR) effect of a change in the value of a particular variable. Specifically, SR effects are those that cause disequilibrium in the system according to well-defined short-term dynamic adjustment processes that push the system back

[7]Note that simultaneity is less likely in our context, as practice managers typically do not choose the workforce based on their preferred level of COC (which is often not even tracked) or other similar performance measures, but rather staffing choices tend to be made on the basis of satisfying patient demand and providing an adequate number of consultation slots.

to its long-run (LR) equilibrium (Granger 1983, Granger et al. 1986). In our case, however, we are more interested in the LR equilibrium relationship between the dependent and independent variables (while accounting for SR shocks to the equilibrium). This requires estimation of the long-run (LR) effects, which capture the impact of a variable on the stable equilibrium (mean or mean with trend) of the dependent variable.

4.6 Autoregressive Distributed Lag Models

4.6.1 Panel ARDL Estimator

To address the limitations described in Section 4.5.3, we consider a family of dynamic non-stationary heterogenous panel data models known as autoregressive distributed lag (ARDL) panel models (Pesaran et al. 1999). These models are denoted by ARDL(J, K) and take the form

$$RC_{pm} = \alpha_p + \sum_{j=1}^{J} \lambda_{pj}^{*} RC_{p(m-j)} + \sum_{k=0}^{K} \boldsymbol{\delta}_{pk}^{*\mathsf{T}} \mathbf{Z}_{p(m-k)} + \epsilon_{pm}\,, \tag{4.3}$$

where J and K specify, respectively, the number of lags of the dependent and independent variables to be included in the model. The set of independent regressors is given by $\mathbf{Z}$, which includes the workload- and workforce-related factors of interest as well as the controls previously specified in $\mathbf{X}$, while the time-varying disturbance term is given by $\epsilon_{pm} \sim \mathcal{N}(0, \sigma^2)$.

This model can be re-expressed in error-correction form by subtracting $RC_{p(m-1)}$ from both sides of the equation, giving (Blackburne III and Frank 2007, Loayza and Ranciere 2004):

$$\Delta RC_{pm} = \alpha_p + \phi_p \left[RC_{p(m-1)} - \boldsymbol{\beta}_p^{\mathsf{T}} \mathbf{Z}_{p(m-1)} \right] + \sum_{j=1}^{J-1} \lambda_{pj} \Delta RC_{p(m-j)} + \sum_{k=0}^{K-1} \boldsymbol{\delta}_{pk}^{\mathsf{T}} \Delta \mathbf{Z}_{p(m-k)} + \epsilon_{pm} \tag{4.4}$$

where $\phi_p = -\left(1 - \sum_{j=1}^{J} \lambda_{pj}^{*}\right)$ and $\boldsymbol{\beta}_p = \sum_{k=0}^{K} \boldsymbol{\delta}_{pk}^{*\mathsf{T}} / \left(1 - \sum_{j=1}^{J} \lambda_{pj}^{*}\right)$.

An important term in this model is the long-run (LR) regression equation $\phi_p \left[RC_{p(m-1)} - \boldsymbol{\beta}_p^{\mathsf{T}} \mathbf{Z}_{p(m-1)} \right]$, which establishes the LR relationship between the dependent variable and the independent variables. The β coefficients thus identify the LR or permanent effect on RC of a change in the independent variables, while the λs and δs capture the short-run (SR) effects of the dependent variable and regressors, respectively. Also important is the coefficient ϕ, which specifies the speed of SR adjustment to the LR equilibrium. Note that the ARDL approach is appropriate so long as there exists a LR relationship among the variables, which requires that ϕ be negative and bounded between -2 and 0 (Samargandi et al. 2015).[8] If ϕ equals zero, then the existence

[8] The ARDL approach is also only valid when the variables are integrated of order zero or one or of a mixture of the two orders. This is an important advantage of the ARDL model, as it makes testing for unit roots unnecessary (Pesaran and Shin 1998). In our case, the Im-Pesaran-Shin panel unit root test, which is typically used for unbalanced panels, provides evidence that all of the variables are integrated of order either zero or one (Im et al. 2003).

of a LR relationship is not supported by the data, while if ϕ falls below -2 or above 0, the process will diverge from rather than converge to the LR equilibrium.

Note that the ARDL technique addresses all of the shortcomings of the panel models outlined in Section 4.5.3 and described further in Section 5.2 of Chapter 5. The addition of multiple lags of both the dependent and independent variables better captures dynamic and temporal dependence in the process and significantly reduces the risk of endogeneity bias, since these lags serve as proxies for other omitted factors. The model also allows for heterogeneity in the slope parameters by allowing for the SR and LR coefficients to be estimated separately for each practice (as specified by the j subscript on the coefficients). Finally, the model is able to distinguish between the LR effects and SR idiosyncratic shocks, which are estimated jointly in the model (Pesaran et al. 1999). This allows us to isolate the permanent impact of the regressors on the LR equilibrium.

Three different estimators can be specified from the general ARDL model in Equation Equation 4.4.[9] First is the mean groups (MG) estimator, a fully heterogeneous model that does not impose any parameter restrictions (Pesaran et al. 1999). Under MG, a separate regression is performed for each practice, and the mean of the LR and SR coefficients are estimated consistently by an unweighted average of the coefficients from the individual regressions. At the other extreme is the dynamic fixed effects (DFE) estimator. This model is based on pooled estimation and assumes homogenous LR and SR coefficients, i.e., the p subscripts on the α, ϕ, β, λ and δ coefficients in Equation Equation 4.4 are dropped. Notice that setting $\phi = 0$, $J = 2$ and $K = 1$ in the DFE model is equivalent to estimating a first-difference (FD) model with one period lagged DV as a control.

The intermediate estimator is the pooled mean groups (PMG) model. Under PMG, the SR coefficients, error correcting speed of adjustment term, regression intercept and error variances are allowed to be heterogeneous across practices, while the LR slope coefficients (i.e., the βs) are restricted to be the same (Pesaran et al. 1999). Consistent SR coefficients are generated by taking the arithmetic mean of the individual practice coefficients (Loayza and Ranciere 2004). The PMG model specification, denoted PMG(J, K), can thus be expressed by replacing the LR regression term in Equation Equation 4.4 with $\phi_p \left[RC_{p(m-1)} - \boldsymbol{\beta}^{\mathsf{T}} \mathbf{Z}_{p(m-1)}\right]$. The parameters of the PMG model are estimated using a maximum likelihood approach (Pesaran et al. 1999), with the lag structure of the model generally determined using a consistent information criterion, such as AIC or BIC.

The PMG estimator is typically preferred in the literature since it is more efficient than the MG estimator when the impact of the regressors on the stable LR equilibrium are homogenous. This also makes PMG particularly appealing in our context because we anticipate that short-term shocks that cause disequilibrium will affect practices differently, yielding practice-specific SR

[9] For all three estimators, the dimensions of N and T are crucial as they should be large enough to apply the dynamic panel technique to ensure unbiasedness of the average estimators (Samargandi et al. 2015).

dynamics. Meanwhile, we have no reason to suspect that the impact of the regressors on the stable LR equilibrium will be heterogenous. We therefore select PMG as our default estimator. (Hausman tests confirm that the PMG is preferred over MG and DFE.)

Following Loayza and Ranciere (2004) and Ahrens (2011), prior to model estimation we eliminate cross-practice common factors by subtracting from each of the variables included in the model their cross-sectional means for each time period. This is an alternative to including time FEs (which cause problems with model convergence in software implementations of panel ARDL), and it ensures consistency of the PMG estimator despite possible cross-sectional dependence, i.e., non-independence of the regression residuals between practices over time caused by, e.g., common omitted factors (Loayza and Ranciere 2004). (Note, however, that results are similar in size and significance if we do not perform this additional de-meaning step.)

4.6.2 Results

Results from the PMG estimation corresponding to the four main regressors are reported in Table 4.3. The top panel of Table 4.3 reports the LR coefficients, whereas the middle panel reports the error correction term ϕ and the SR coefficients. The bottom panel describes the structure of the controls as well as the value of the Akaike's information criteria (AIC) corresponding to each model. Estimation is performed using the `xtpmg` routine in Stata 16.

Table 4.3 PMG estimates of the long- and short-run effects.

	(1)	(2)	(3)
Long-Run			
$zPracticePop$	-0.080***	-0.058***	-0.074***
	(0.009)	(0.009)	(0.009)
$zConsPerPat$	-0.024***	-0.019***	-0.037***
	(0.002)	(0.002)	(0.002)
$zDaysPerEstGP$	0.050***	0.038***	0.049***
	(0.002)	(0.002)	(0.002)
$zShareEstGP$	0.050***	0.040***	0.043***
	(0.002)	(0.001)	(0.002)
Short-Run			
EC term (ϕ_p)	-0.249***	-0.350***	-0.229***
	(0.008)	(0.010)	(0.008)
1st order lag of ΔRC	-0.166***	-0.139***	-0.141***
	(0.007)	(0.007)	(0.006)
$\Delta zPracticePop$	-0.014	0.000	-0.020
	(0.056)	(0.056)	(0.057)
$\Delta zConsPerPat$	-0.001	0.000	0.006**
	(0.002)	(0.002)	(0.002)
$\Delta zDaysPerEstGP$	0.005***	0.005***	0.003**
	(0.001)	(0.001)	(0.001)
$\Delta zShareEstGP$	0.035***	0.030***	0.038***
	(0.002)	(0.002)	(0.002)
Practice FE	Yes	Yes	Yes
Controls	None	SR Only	SR & LR
Observations	29,643	29,643	29,643
AIC	-3.85	-3.91	-3.92

Notes: $^{+}\,p < 0.10$, $^{*}\,p < 0.05$, $^{**}\,p < 0.01$, $^{***}\,p < 0.001$; Standard errors in parentheses; EC refers to the error-correcting speed of adjustment term.

Columns (1)–(3) represent different configurations of control structures in the PMG(J, K) models. Column (1) corresponds to the model without controls, while column (2) includes controls as fixed regressors (i.e., appearing only in the SR equation). Lastly, column (3) gives results from the full model in which controls are included as dynamic regressors and hence appear in both the LR and SR equations. We follow the literature and select the lag structure by allowing $J, K \in \{1, 2, 3, 4\}$ and estimating all 16 permutations of each model, then selecting for each column the lag structure that produces the lowest AIC. In each case, this corresponds to a PMG$(2, 1)$ model, i.e., the model specified by the equation:

$$\Delta RC_{pm} = \alpha_p + \phi_p \left[RC_{p(m-1)} - \boldsymbol{\beta}^{\mathsf{T}} \mathbf{Z}_{p(m-1)} \right] + \lambda_p \Delta RC_{p(m-1)} + \boldsymbol{\delta}_p^{\mathsf{T}} \Delta \mathbf{Z}_{pm} + \epsilon_{pm} \,. \tag{4.5}$$

As the results are similar across models and since the model in column (3) has lowest AIC, we

proceed to interpret the full model. First, observe that the value of the error correcting speed of adjustment term (i.e., ϕ) takes value -0.229, indicating the existence of a LR relationship and validating the use of the ARDL approach. Next, we note that with this model we achieve a within-practice R^2 of 62.5%, considerably higher than the 34.8% value estimated with full FE model (see column (6) of Table 4.2). Excluding the four main independent variables from the ARDL model reduces the R^2 value to 41.7%, indicating that the four workload and workforce factors alone can explain 35.7% (=$(62.5 - 41.7)/(100 - 41.7)$) of the residual variation in RC.

Turning to the coefficient estimates, recall that our primary focus is on the LR coefficients, which capture the permanent effect of a change in the independent variables on the LR equilibrium. Starting with the workload-related factors and all else remaining equal, a 1σ increase in *PracticePop* or *ConsPerPat* leads to a 7.4 p.p. or 3.7 p.p. reduction in RC, respectively. As for the workforce-related factors, we find a 1σ decrease in *DaysPerEstGP* or *ShareEstGP* causes RC to decrease by 4.9 p.p or 4.3 p.p., respectively. All effects are significant at the 0.1% significance level.

To improve interpretation of the results, in Table 4.4 we use the between-practice standard deviation (BPSD) from Table 4.1 to examine how the average RC rate is expected to vary across practices based on practice characteristics. For example, comparing a large practice with a small practice, where the former has a *PracticePop* two BPSDs above the mean and the latter two BPSDs below the mean, shows that the larger practice will have a RC rate 50.5% (= $(0.290 - 0.586)/0.586$) lower than that of the smaller practice, with RC reduced by 29.6 p.p. Meanwhile, a practice that relies more on part-timers or unestablished GPs (with *DaysPerEstGP* or *ShareEstGPs* two BPSDs below the mean) will only be able to match patients with their regular provider 38.6% of the time. This is 24.0% lower than a practice that uses predominantly full-time salaried workers or established GPs, which matches patients with their preferred provider 50.8% of the time. These factors are therefore highly consequential in explaining variation between practices in their ability to provide RC.

Table 4.4 Average variation in relational continuity across practices, by practice characteristics.

Variable	-2σ	-1σ	$+0\sigma$	$+1\sigma$	$+2\sigma$
Smaller vs. larger practice population	0.586	0.512	0.438	0.364	0.290
Lower vs. higher consultation rate per patient	0.500	0.469	0.438	0.407	0.376
Fewer vs. more days worked per month by est. GPs	0.386	0.403	0.438	0.473	0.508
Lower vs. higher dependence on est. GPs	0.386	0.403	0.438	0.473	0.508

Notes: This comparison is based on the variation that exists between practices rather than total variation. For example, the $+1\sigma$ column shows the impact of a one unit increase in the between variation for the relevant variable, so for *zConsPerPat* this would equal the estimated coefficient from Table 4.3, -0.037, multiplied by the between variation reported in Table 4.1, i.e., 0.84.

4.6.3 Explaining the Trend in Relational Continuity

Next, we investigate which factors are most important in explaining the decline in RC over time. In Table 4.5 we report the change in the value of each of the main variables over the observation period. The change is calculated by estimating an OLS regression for each variable of the form $y_{pm} = \alpha_p + \mu t_{pm} + \boldsymbol{\nu}^{\mathsf{T}}\mathbf{M}_{pm} + \epsilon_{pm}$ using a weighted estimation, where the weights are proportional to $\left|C^{+}_{pm}\right|$.[10] The trend term, t_{pm}, is a variable that takes starting value zero in January 2008 and increases in value by one unit for every month into our study. The coefficient μ thus captures the average monthly change in the dependent variable. Meanwhile, α_p is a practice FE that controls for the fact that the starting values of the dependent variables may differ across practices and also for the fact that some practices drop out of the sample during the observation period, influencing μ. Also included is a vector of month-of-the-year dummies, $\mathbf{M}_{pm}$, that account for seasonality.

Table 4.5 Trend in key variables.

Variable	$t = 0$ [a]	$t = 119$ [a]	Δ	$\Delta \times \beta$	% trend explained
RC	0.48	0.36	-0.126p.p. [b]	–	–
PracticePop [c]	7.86	8.64	0.199σ	-0.0148p.p	11.8
ConsPerPat	0.29	0.30	0.118σ	-0.0044p.p	3.5
DaysPerEstGP	13.41	11.71	-0.525σ	-0.0256p.p	20.4
ShareEstGP	0.81	0.74	-0.382σ	-0.0162p.p	12.9

Notes: [a] Estimated using the OLS model described in the first paragraph of Section 4.6.3, then taking the unweighted average of the predicted values of the dependent variable across all practices at time $t = 0$ (i.e., month 0) and $t = 119$ (i.e., month 119); [b] p.p. indicates a percentage point change; [c] *PracticePop* reported in thousands.

Using the above approach to estimate μ, in the Δ column of Table 4.5 we report the overall change in each variable over the sample period, which is equal to 119μ, where 119 is the number of months between the first and last month in our study. Multiplying this value by the estimated changes associated with a 1σ increase in each variable (taken from column (3) of Table 4.3) allows us to identify the contribution of each of the workload and workforce factors to the total reduction in RC, which is given in the $\Delta \times \beta$ column of Table 4.5. This shows, for example, that the 0.199σ increase in the average size of the patient list can explain $\simeq$11.8% of the 0.126 p.p. reduction in RC ($= (0.199 \times -0.074)/-0.126$). On the other hand, the 0.525σ shift to a more fragmented part-time workforce over the ten-year time horizon explains $\simeq$20.4% of the reduction in RC, with an increase in reliance on the non-established workforce by 0.382σ explaining a further $\simeq$12.9% reduction. Together, these three factors alone can thus explain nearly half ($\simeq$45.0%) of the total reduction in RC over the ten-year observation period. Adding to this, the increase in

[10] This weighting gives higher relative importance to practices that treat more patients and months in which more patients are seen, so it better captures the trend in the population than a simple unweighted average across practice-months. Our results are, however, nearly identical without the weighting scheme.

the number of consultations per patient by 0.118σ over the sample period results in a decrease in RC by a further 0.004 p.p. This is of relatively lower importance operationally and increases the reduction in RC explained by an additional $\simeq$3.5%, to 48.5%.

4.6.4 Robustness

We have conducted a range of robustness checks to ensure that our results and insights are not confined to the specifications presented in the main manuscript. First, as noted in Footnote 4, we have re-estimated the results using the subsample of 79 practices continuously present in the CPRD dataset. Second, we have reproduced our findings using different definitions of a full working day, changing the threshold to require a minimum of either 5 or 15 consultations, as mentioned in Section 4.4.2.3. Third, we have repeated the analysis using a different approach to handling ties when identifying the regular GP. Specifically, when a patient has had the same number of appointments with two or more providers over the past two years, we instead break the tie by assigning the GP who the patient saw most recently as the regular GP. All results are reported in Section 5.3-5.5 of Chapter 5, with all findings consistent.

In addition, while the lag structure in Section 4.6.2 was chosen to minimize the AIC, an advantage of the ARDL model is that when enough lags are included in the SR equation, the model provides consistent coefficients despite the possible presence of endogeneity (Pesaran et al. 1999, Pesaran and Shin 1998). For this reason, to check our results against possible endogeneity bias, we have also estimated models with different lag structures, specifically with $J, K = 1, 2, 3$ and 4, respectively. Results are consistent and reported in Section 5.6 of Chapter 5, indicating that our findings are robust against the presence of potential omitted variables.

4.7 Managerial Implications and Conclusions

Primary care providers around the world are facing the dual challenge of managing an increasing demand for healthcare resources and contending with changes in the size and composition of the workforce. A common response has been to replace or augment permanent employees with temporary workers, counter burnout by allowing staff to switch to part-time work patterns, and manage the overall workload by adopting more flexible pooled scheduling practices. In this paper, we have demonstrated that these responses – which may be in some instances unavoidable – have also made it significantly harder for patients to access their providers of choice, causing a deterioration in RC. As the first paper to demonstrate the important role of operational factors such as workforce composition and workload on RC in the primary care setting, this study has a number of immediate implications for practice.

First, this helps GP practice managers be aware of what factors are important for providing continuity and identifying the root causes of low continuity at their practice.

Policymakers should also consider the implications on continuity when setting financial targets for providing speedy access at the expense of continuity, and while deciding policies relating to the workforce. Additionally, policymakers must improve the attractiveness of full-time established employment if they wish to preserve RC. This is a view that is starting to gain traction, with *"many practices now report[ing] that a shift to reliance on locums is undermining service continuity and stable team working"* and growing recognition that it is *"in the interests of GPs and practices to improve the relative attractiveness of partner and salaried positions versus a shift to a more unstable and short term workforce"* (NHS England 2016b, p. 23). Mitigating the adverse effects of workload for all practice staff is one step towards achieving this shift, as higher workload not only causes a direct reduction in RC but also creates conditions (e.g., stress and burnout) that may lead to workforce fragmentation. While it is not within the scope of this work to prescribe precisely how to improve working conditions,[11] this paper does help provide the impetus to do so. In particular, we contribute the first piece of empirical evidence that maintaining an established workforce and making the work attractive enough to keep GPs in full-time employment will have a significant impact on RC. With RC already known to improve patient outcomes, this finding is not only of operational interest, but it also has direct clinical implications.

Importantly for practice, our results suggest that contrary to the wide-held belief that the decline in RC is primary driven by increases in workload factors due to population growth and the intensity of resource use by patients, instead, workforce factors appear to explain most of the fall in our study context. This has important implications for practice managers. For example, a recent and controversial trend has been for GP practices to seek government approval to suspend new patient signups in order to manage increases in workload (NHS England 2016a). Our results suggest that while this may help to ease immediate workload pressures, this should not be seen as a solution that will have a significant bearing on RC provision in the long-term.

Second, when workload and workforce changes are unavoidable, GP practice managers should be aware of the potential adverse effects on RC provision and adopt proactive strategies to minimize these effects. While this paper does not explore patient-specific moderators, one approach indicated by the literature is to prioritize RC for those patients who will benefit from it the most: for example, older patients or those with chronic conditions (Kajaria-Montag et al. 2020). These patients might, for instance, be allocated to full-time established GPs or might be given priority access to their preferred providers on arrival.

Third, our results can help explain why some practices are less able to provide RC than others. In particular, we observe, based on both our own interactions with practice managers and on statements from professional bodies representing primary care providers (Jeffers and Baker 2016), that RC is widely recognized as a cornerstone of the delivery model used in general

[11] We note that proposals to increase the attractiveness of full-time GP positions include providing childcare arrangements, increasing patient-facing time, adopting helpful technologies (e.g., telemedicine), improving retirement benefits, and introducing schemes to reduce burnout and workload-related stress.

practice. However, providers are often unaware of the root causes of low rates of RC within their own practices. Our paper can help practice managers to answer this question by characterizing the impact of practice characteristics on RC (see, e.g., Table 4.4), thus enabling them to design targeted interventions to improve RC provision. We also note that the decline in RC is an issue that patients care about: the majority of patients value interpersonal care continuity, yet many report that they are unable to see their preferred provider (Aboulghate et al. 2012). Comparing the operational characteristics of practices can thus help patients, especially those who value RC most highly, to choose the practice that is best for them.

We note also that although this work is focused on the primary care context, it provides valuable insights and methodologies that may also prove useful in the study of other knowledge-intensive or relationship-based services (e.g., banking, legal, consultancy, etc.). In such settings, knowledge accumulation and learning that occurs through repeated interactions can help to improve service outcomes, quality, and efficiency (Henderson and Tookes 2012, Taylor and Plambeck 2007, Huckman and Pisano 2006). However, exploiting such benefits requires firms to be able to match customers with servers on an ad-hoc basis as and when demand materializes. We demonstrate the critical role that staff scheduling and the use of in-house (rather than external) service providers can play in facilitating these repeated interactions. Moreover, we provide an estimation approach that can be adopted in other contexts for identifying the relative importance of different factors in explaining a firm's ability to provide customer-server continuity.

With respect to methodology, we introduce to the OM community the panel ARDL modeling framework which, until now, has primary been used by researchers in the field of macroeconomics. As highlighted in Section 4.6, ARDL models have many advantages over traditional panel methods such as the FE and FD estimators. In particular, in a dynamic setting panel ARDL methods enable direct estimation of the long-term or permanent impact of the independent variables on the LR equilibrium while also addressing endogeneity bias through inclusion of firm-specific fixed effects and lags of the dependent and independent variables in the SR estimation equation. We anticipate that this novel approach can be easily extended to other empirical OM contexts, both within and outside of the healthcare domain, especially as researchers are increasingly gaining access to large macro panels (i.e., large N and large T).

Finally, we note that while the focus of this paper is on robustly identifying the size and relative importance of the workload- and workforce-related factors under study, our work may be extended in a number of ways. First, while we are able to explain a significant percentage of the long-term decline in RC over the study period (and $\simeq$63% of the within-practice variation), a reasonable proportion of the decline remains unexplained. This indicates there may be other important factors unobservable to us (e.g., waiting time for an appointment) that are available in other datasets and which could contribute additional insights. Second, we focus on establishing the aggregate effects in this study, but further research may look to identify the types of changes that occur in practice (e.g., to scheduling practices) as workload and workforce composition varies, allowing for a better understanding of how exactly our variables of interest lead

to a fall in RC. Third, not all GP practices are affected equally by the factors under study, and some appear more resilient than others to the increase in workload and changes in workforce composition. Further work may thus look to identify moderators (e.g., use of technology) that can help to increase levels of RC despite these challenging trends.

Overall, as the first paper to demonstrate empirically the importance of operational factors in driving variation in RC between practices and over time, this work provides important insights for practice and provokes a range of follow-up questions that might be pursued in future research.

References

Aboulghate A, Abel G, Elliott MN, Parker RA, Campbell J, Lyratzopoulos G, Roland M (2012) Do english patients want continuity of care, and do they receive it? *British Journal of General Practice* 62(601):e567–e575.

Ahrens A (2011) Do labour market institutions influence consumers' saving intentions? aggregate evidence from europe. Technical report, IAAEG Discussion Paper Series.

Ahuja V, Alvarez CA, Staats BR (2020a) How continuity in service impacts variability: Evidence from a primary care setting, SMU Cox School of Business Research Paper No. 19-13.

Ahuja V, Alvarez CA, Staats BR (2020b) Maintaining continuity in service: An empirical examination of primary care physicians. *Manufacturing & Service Operations Management* .

Amjad H, Carmichael D, Austin AM, Chang CH, Bynum JP (2016) Continuity of care and health care utilization in older adults with dementia in fee-for-service medicare. *JAMA internal medicine* 176(9):1371–1378.

Anand KS, Paç MF, Veeraraghavan S (2011) Quality-speed conundrum: Trade-offs in customer-intensive services. *Management Science* 57(1):40–56, ISSN 00251909.

Baird B, Holmes J (2019) Why can't I get a doctor's appointment? Technical report, The King's Fund, URL `https://www.kingsfund.org.uk/publications/solving-issue-gp-access`.

Barker I, Steventon A, Deeny SR (2017) Association between continuity of care in general practice and hospital admissions for ambulatory care sensitive conditions: Cross sectional study of routinely collected, person level data. *BMJ (Online)* 356, ISSN 17561833.

Beech J, Bottery S, Charlesworth A, Evans H, Gershlick B, Hemmings N, Imison C, Kahtan P, McKenna H, Murray R, et al. (2019) Closing the gap. *Key areas for action on the health and care workforce. London: The Health Foundation/Nuffield Trust/The King's Fund* .

Benjaafar S (1995) Performance bounds for the effectiveness of pooling in multi-processing systems. *European Journal of Operational Research* 87(2):375–388.

Blackburne III EF, Frank MW (2007) Estimation of nonstationary heterogeneous panels. *The Stata Journal* 7(2):197–208.

Bobroske K, Freeman M, Huan L, Cattrell A, Scholtes S (2020) Curbing the Opioid Epidemic at its Root: The Effect of Provider Discordance after Opioid Initiation. *SSRN Electronic Journal* .

Bostock N (2018) More than 1,100 GP practices have closed or merged under NHS England.

Bostock N (2019a) Less than three in 10 GPs work full time, official data show.

Bostock N (2019b) Number of GP practices in England falls below 7,000. *GPonline* .

Boyle S, Appleby J, Harrison A (2010) A rapid view of access to care. Technical report, The Kings Fund.

Cachon G, Terwiesch C (2011) *Matching Supply with Demand* (McGraw-Hill Professional), 3 edition.

Campos N, Kinoshita Y (2008) Foreign direct investment and structural reforms: Panel evidence from eastern europe and latin america. *IMF Staff Papers* .

Canales R, Greenberg J (2016) A matter of (relational) style: Loan officer consistency and exchange continuity in microfinance. *Management Science* 62(4):1202–1224.

Centre for Workforce Intelligence (2014) In-depth review of the general practitioner workforce. *Centre for Workforce Intelligence* (July).

Cho KH, Kim YS, Nam CM, Kim TH, Kim SJ, Han KT, Park EC (2015) The association between continuity of care and all-cause mortality in patients with newly diagnosed obstructive pulmonary disease: a population-based retrospective cohort study, 2005-2012. *PloS one* 10(11).

Dalen JE, Ryan KJ, Alpert JS (2017) Where Have the Generalists Gone? They Became Specialists, Then Subspecialists. *American Journal of Medicine* 130(7):766–768, ISSN 15557162.

Donnelly L (2018) Average GP now works 3.5 days a week - and just one in 20 trainees plans to do the job full-time. *The Telegraph* .

Doran N, Fox F, Rodham K, Taylor G, Harris M (2016) Lost to the NHS: A mixed methods study of why GPS leave practice early in England. *British Journal of General Practice* 66(643):e128–e134.

Dossa AR, Moisan J, Guénette L, Lauzier S, Grégoire JP (2017) Association between interpersonal continuity of care and medication adherence in type 2 diabetes: an observational cohort study. *CMAJ open* 5(2):E359.

Drury A, Payne S, Brady AM (2020) Identifying associations between quality of life outcomes and healthcare-related variables among colorectal cancer survivors: A cross-sectional survey study. *International journal of nursing studies* 101:103434.

Gault B (2019) Average GP waiting times exceed two weeks for first time ever.

General Medical Council (2018) What our data tells us about GPs working for the NHS in England and Scotland. Working paper, General Medical Council.

Graham Clews (2013) Exclusive: Most GPs say their job has become more stressful — GPonline. URL `https://www.gponline.com/exclusive-gps-say-job-become-stressful/article/1207601`.

Granger CW (1983) *Co-integrated variables and error-correcting models*. Ph.D. thesis, UCSD Discussion Paper 83-13.

Granger CWJ, et al. (1986) Developments in the study of cointegrated economic variables. *Oxford Bulletin of economics and statistics* (Citeseer).

Grembowski D, Paschane D, Diehr P, Katon W, Martin D, Patrick DL (2005) Managed care, physician job satisfaction, and the quality of primary care. *Journal of General Internal Medicine* 20(3):271–277.

Haggerty JL, Reid RJ, Freeman GK, Starfield BH, Adair CE, McKendry R (2003) Continuity of care: a multidisciplinary review. *Bmj* 327(7425):1219–1221.

Hallvik SE, Geissert P, Wakeland W, Hildebran C, Carson J, O'kane N, Deyo RA (2018) Opioid-prescribing continuity and risky opioid prescriptions. *The Annals of Family Medicine* 16(5):440–442.

Heath I (1995) Fortnightly Review: Commentary: The perils of checklist medicine. *Bmj* 311(7001):373, ISSN 14685833, URL `http://dx.doi.org/10.1136/bmj.311.7001.373`.

Heiser S (2019) New Findings Confirm Predictions on Physician Shortage. *AAMC News* 1–3.

Henderson BJ, Tookes H (2012) Do investment banks' relationships with investors impact pricing? the case of convertible bond issues. *Management Science* 58(12):2272–2291.

Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, Van Staa T, Smeeth L (2015) Data Resource Profile Data Resource Profile: Clinical Practice Research Datalink (CPRD). *International Journal of Epidemiology* 827–836.

Hobbs FR, Bankhead C, Mukhtar T, Stevens S, Perera-Salazar R, Holt T, Salisbury C, et al. (2016) Clinical workload in uk primary care: a retrospective analysis of 100 million consultations in england, 2007–14. *The Lancet* 387(10035):2323–2330.

Huckman RS, Pisano GP (2006) The firm specificity of individual performance: Evidence from cardiac surgery. *Management Science* 52(4):473–488.

Huntley A, Lasserson D, Wye L, Morris R, Checkland K, England H, Salisbury C, Purdy S (2014) Which features of primary care affect unscheduled secondary care use? a systematic review. *BMJ open* 4(5):e004746.

Im KS, Pesaran MH, Shin Y (2003) Testing for unit roots in heterogeneous panels. *Journal of econometrics* 115(1):53–74.

Institute for Government (2019) General practice — The Institute for Government.

Jeffers H, Baker M (2016) Continuity of care: still important in modern-day general practice. *British Journal of General Practice* 66(649):396–397.

Jouini O, Dallery Y, Nait-Abdallah R (2008) Analysis of the impact of team-based organizations in call center management. *Management Science* 54(2):400–414.

Kajaria-Montag H, Freeman M, Scholtes S (2020) The impact of relational continuity on primary care operations, Working Paper.

Katz DA, McCoy KD, Vaughan-Sarrazin MS (2015) Does greater continuity of veterans administration primary care reduce emergency department visits and hospitalization in older veterans? *Journal of the American Geriatrics Society* 63(12):2510–2518.

Kings Fund (2015) Long-term conditions and multi-morbidity — The King's Fund. Technical report, KingFund.

Kristjansson E, Hogg W, Dahrouge S, Tuna M, Mayo-Bruinsma L, Gebremichael G (2013) Predictors of relational continuity in primary care: patient, provider and practice factors. *BMC family practice* 14(1):72.

Legraien L (2019) Four in 10 patients offered same-day GP appointments.

Levene LS, Baker R, Walker N, Williams C, Wilson A, Bankart J (2018) Predicting declines in perceived relationship continuity using practice deprivation scores: a longitudinal study in primary care. *Br J Gen Pract* 68(671):e420–e426.

Loayza N, Ranciere R (2004) *Financial development, financial fragility, and growth* (The World Bank).

Maarsingh OR, Henry Y, van de Ven PM, Deeg DJ (2016) Continuity of care in primary care and association with survival in older people: a 17-year prospective cohort study. *Br J Gen Pract* 66(649):e531–e539.

Matthews-King A (2015) GP practices' locum use surges 20% in a year.

NHS Digital (2017) Patients Registered at a GP Practice December 2017 - NHS Digital. Technical report, NHS Digital.

NHS England (2012) Attribution dataset gp registered populations scaled to ons population estimates?2011. *London: Health and Social Care Information Centre* .

NHS England (2016a) Commissioner Guidelines for Responding to Requests from Practices to Temporarily Suspend Patient Registration. Technical report, NHS England.

NHS England (2016b) General practice forward view. Technical report, NHS England.

NHS England (2018) NHS england. Technical report, NHS England.

NHS Improvement (2011) It's Your Practice A patient guide to GP services. Technical report, NHS England, URL `www.rcgp.org.uk`.

Nyweide DJ, Anthony DL, Bynum JP, Strawderman RL, Weeks WB, Casalino LP, Fisher ES (2013) Continuity of care and the risk of preventable hospitalization in older adults. *JAMA internal medicine* 173(20):1879–1885.

Office for National Statistics (2019) Estimates of the population for the UK, England and Wales, Scotland and Northern Ireland - Office for National Statistics.

Palmer B (2019) Is the number of gps falling across the uk? *blog, Nuffield Trust. https://www. nuffield-trust. org. uk/news-item/is-the-number-of-gps-falling-across-the-uk* .

Payne RA, Mendonca SC, Elliott MN, Saunders CL, Edwards DA, Marshall M, Roland M (2020) Development and validation of the cambridge multimorbidity score. *CMAJ* 192(5):E107–E114.

Pesaran MH, Shin Y (1998) An autoregressive distributed-lag modelling approach to cointegration analysis. *Econometric Society Monographs* 31:371–413.

Pesaran MH, Shin Y, Smith RP (1999) Pooled mean group estimation of dynamic heterogeneous panels. *Journal of the american statistical association* 94(446):621–634.

Queenan C, Cameron K, Snell A, Smalley J, Joglekar N (2019) Patient heal thyself: reducing hospital readmissions with technology-enabled continuity of care and patient activation. *Production and Operations Management* 28(11):2841–2853.

Ride J, Kasteridis P, Gutacker N, Doran T, Rice N, Gravelle H, Kendrick T, Mason A, Goddard M, Siddiqi N, et al. (2019) Impact of family practice continuity of care on unplanned hospital use for people with serious mental illness. *Health services research* 54(6):1316–1325.

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2012) Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research* 60(5):1080–1097.

Salisbury C, Sampson F, Ridd M, Montgomery AA (2009) How should continuity of care in primary health care be assessed? *British Journal of General Practice* 59(561):276–282, ISSN 09601643.

Samargandi N, Fidrmuc J, Ghosh S (2015) Is the relationship between financial development and economic growth monotonic? evidence from a sample of middle-income countries. *World development* 68:66–81.

Senot C (2019) Continuity of care and risk of readmission: An investigation into the healthcare journey of heart failure patients. *Production and Operations Management* 28(8):2008–2030.

Smith DR, Whitt W (1981) Resource sharing for efficiency in traffic systems. *Bell System Technical Journal* 60(1):39–55.

Song H, Tucker AL, Murrell KL (2015) The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science* 61(12):3032–3053.

Tammes P, Payne RA, Salisbury C, Chalder M, Purdy S, Morris RW (2019) The impact of a named gp scheme on continuity of care and emergency hospital admission: a cohort study among older patients in england, 2012–2016. *BMJ open* 9(9):e029103.

Tammes P, Purdy S, Salisbury C, Mackichan F, Lasserson D, Morris RW (2017) Continuity of primary care and emergency hospital admissions among older patients in England. *Annals of Family Medicine* 15(6):515–522, ISSN 15441717.

Taylor TA, Plambeck EL (2007) Supply chain relationships and contracts: The impact of repeated interaction on capacity investment and procurement. *Management science* 53(10):1577–1593.

The Physicians Foundation (2018) 2018 Survey of America's physicians. Technical report, The Physicians Foundation.

Verbeek M, Nijman T (1992) Testing for selectivity bias in panel data models. *International Economic Review* 681–703.

Wang J, Zhou YP (2018) Impact of queue configuration on service time: Evidence from a supermarket. *Management Science* 64(7):3055–3075.

Ye T, Sun X, Tang W, Miao Y, Zhang Y, Zhang L (2016) Effect of continuity of care on health-related quality of life in adult patients with hypertension: a cohort study in china. *BMC health services research* 16(1):674.

Chapter 5

The Operational Determinants of Relational Continuity of Care – Further investigations

5.1 Introduction

This chapter includes supporting material designed to go with the analysis presented in Chapter 4 and therefore is not meant to be read in isolation but rather serves as material that provides more background and support for the methods, techniques and results presented, as well as helps to confirm the robustness of the main findings. In Section 5.2, we discuss the limitations of the fixed effects estimator in detail. In Section 5.3, we repeat the analysis with a balanced panel of 79 practices. In Sections 5.4, 5.5 and 5.6, we consider alternative cutoffs for the workforce variables, alternative tie-breaking methods for assignment of the regular doctor, and alternative lag structures for the PMG model, respectively. In Section 5.7, we consider a short practice-year panel and repeat the analysis using the Generalised Method of Moments estimator. Finally, in Section 5.8, we consider an alternative framing that only looks at various aspects of workforce fragmentation as independent variables.

5.2 Limitations of the fixed effects estimator

The FE estimator described provides preliminary evidence that all of the factors being studied have an impact on RC. However, there are a number of limitations with the FE estimator that we discuss below and which drive us to adopt an alternative modeling approach in Section 4.6 of Chapter 4.

Non-stationarity.

One concern with macro (i.e., large N and large T) panels is non-stationarity, which can lead to spurious regression estimates (Baltagi 2008). Non-stationarity is typically dealt with by replacing the FE estimator with the first difference (FD) estimator. The FD model takes the first differences of both the dependent and independent variables and in doing so removes the

incidental parameters (i.e., the α_p terms) as well as any time-invariant omitted variable from the error term. The coefficients of the FD estimator have the same interpretation as those of the FE estimator, and they are reported in column (1) in Table 5.1. Findings remain consistent in direction, though the effects of the practice size and consultation rate become statistically indistinguishable from zero.

Dynamic misspecification and serial correlation.

The standard FE and FD estimators assume serially uncorrelated disturbances. If this assumption is violated in a static regression, then serial correlation has consequences similar to heteroskedasticity (which can be addressed, for instance, by estimating robust standard errors). However, evidence of serial correlation may also be a sign of misspecification of the underlying model, e.g., if the true model is dynamic but is wrongly assumed to be static (Balestra 1982). A model is said to be dynamic if history matters, i.e., if the dependent variable is influenced not only by the current value of the independent variable(s), but also by values of the independent variable(s) in the past. If these dynamics are present but not sufficiently captured, coefficient estimates may be biased (Campos and Kinoshita 2008).

Testing for serial correlation in our FE/FD models using a procedure proposed by Wooldridge (2002) provides evidence to suggest that, indeed, the within-group error terms are serially correlated. One approach to (at least partially) address this issue and that of omitted variable bias more generally is to include the lag of the dependent variable on the right-hand side of the regression. (Note that in short (i.e., small T) panels, introducing the lag on the DV on the RHS of a fixed effects model can lead to a bias of order $1/T$ as $N \to \infty$, which is referred to as Nickell's bias (Nickell 1981). In macro panels like ours where T is relatively large, this is not a major concern.) In effect, the lag of the DV accounts for dynamic and temporal dependence in the process as well as serving as a proxy variable to capture other unobserved factors (Wooldridge 2002, Gokpinar et al. 2010). The FD model with first lag of the DV as a control can be written

$$\begin{aligned} \Delta RC_{pm} &= \gamma_m + \beta_0 \Delta RC_{p(m-1)} + \beta_1 \Delta PracticePop_{pm} + \beta_2 \Delta ConsPerPat_{pm} \\ &+ \beta_3 \Delta DaysPerEstGP_{pm} + \beta_4 \Delta ShareEstGPs_{pm} \\ &+ \boldsymbol{\beta}_5^{\mathsf{T}} \Delta \mathbf{X}_{pm} + \Delta \epsilon_{pm} \,. \end{aligned} \tag{5.1}$$

Table 5.1 First difference regression results.

	(1)	(2)
$zPracticePop$	-0.040	-0.045
	(0.025)	(0.028)
$zConsPerPat$	0.001	0.001
	(0.003)	(0.003)
$zDaysPerEstGP$	0.016***	0.012***
	(0.002)	(0.002)
$zShareEstGP$	0.051***	0.048***
	(0.003)	(0.003)
1st order lag of ΔRC		-0.262***
		(0.011)
Constant	-0.001***	-0.001***
	(0.000)	(0.000)
Controls	Yes	Yes
Time FE	Yes	Yes
Month of Year	Yes	Yes
Observations	29963	29643
R^2	0.303	0.366

Notes: $^{+}\, p < 0.10$, $^{*}\, p < 0.05$, $^{**}\, p < 0.01$, $^{***}\, p < 0.001$; Standard errors, clustered by practice, in parentheses; R^2 specifies the residual variance in RC explained after accounting for practice FE, with practice FEs alone explaining 75% of the variation; The number of observations is reduced by 320 (i.e., one per practice) in column (2) since the first observation is lost for each practice in the first differencing process, and the number of observations is further reduced in column (3) as the first lag of the first differenced dependent variable is included in the regressor on the right-hand side.

Estimating this model and reporting results in column (2) of Table 5.1, we find that the one-period lag of the DV is highly predictive, taking value -0.262 (p-value < 0.001). Meanwhile, the main findings remain unchanged from those in column (1), indicating that there are unlikely to be important omitted variables that are correlated with the main regressors of interest. Note, however, that this is the simplest dynamic panel model that can be specified, including only the first lag of the DV on the RHS and no lags of the IVs – a limitation we address in Section 4.6.

Slope heterogeneity.

While the traditional FE/FD models assume heterogeneity in the intercept term across different practices, they also assume homogeneity in the slope parameters across practices. When time-varying factors affect the practices differently, this assumption will be violated, resulting in inconsistent parameter estimates (Ul Haque et al. 2005). This may be especially problematic in dynamic models, such as the one specified in Equation Equation 5.1, in which assuming homogeneity of the coefficient(s) of the lagged DV can lead to serious bias (Samargandi et al. 2015). In our context, we have reason to suspect that slope parameters may vary across practices.

For example, practices vary considerably in their scheduling practices, as discussed in Section 2.4.1. Additionally, different practices in different regions have been testing new models of care, e.g., forming networks, developing off-hour services, and merging into so-called super-partnerships (Smith et al. 2013). These differences in practice management and structure may affect the extent to which RC is affected by changes in the various workload and workforce factors over time. We discuss a model to overcome this limitation in Section 4.6.

Short-run versus long-run effects.

As noted when discussing the issue of dynamic misspecification above, the effect on RC of the independent variables may not only have an immediate effect, but it can also affect future values. However, in the models specified so far, the estimated coefficients on the current value of the independent variables measure only the short-run (SR) or impact effect of these variables on RC. The long-run (LR) effect, which takes account of both the current and lagged effects, will often be larger (Baltagi and Griffin 1984). To see this, take a simple dynamic model, e.g., of the form specified in Equation Equation 5.1, in which $y_t = \beta_0 + \beta_x x_t + \beta_y y_{t-1} + \epsilon_t$ where $|\beta_y| < 1$. The LR effect in this model can then be approximated by $\beta_x/(1-\beta_y)$. We can also estimate the LR and SR relationships explicitly by subtracting y_{t-1} from both sides and rewriting in error correction form, i.e., $\Delta y_t = \beta_0 + \phi(y_{t-1} - \theta x_{t-1}) + \beta_x \Delta x_t + \epsilon_t$, where $\phi = (\beta_y - 1)$. Here, the SR effect is estimated by β_x and the LR effect by θ via maximum likelihood (Reed and Zhu 2017).

We discuss the difference in the interpretation of the SR and LR effects further in Section 4.5.3 of Chapter 4.

5.3 Subsample of 79 practices

Table 5.2 reports results from the subset analysis using only those practices that are continuously present in the CPRD dataset during the study period, giving us a balanced panel dataset of 79 practices over ten years, as described in Section 4.4.1.2 of Chapter 4. Results are consistent with those reported in Chapter 4. (Note that since the ARDL(2,1) models include the lag of the first difference of the dependent variable on the right-hand side of the equation, this means that the first two observations of each practice are lost. This leaves 118 observations ($= 10 \times 12 - 2$) per practice, which when multiplied by the number of practices (79) gives a total of 9,322 observations.)

Table 5.2 PMG estimates of the long- and short-run effects on RC – Uses a balanced panel of the 79 practices present in all time periods.

	(1)	(2)	(3)
Long-Run			
$zPracticePop$	-0.080***	-0.057***	-0.066***
	(0.010)	(0.009)	(0.010)
$zConsPerPat$	-0.017***	-0.011***	-0.032***
	(0.002)	(0.002)	(0.002)
$zDaysPerEstGP$	0.040***	0.029***	0.040***
	(0.002)	(0.001)	(0.002)
$zShareEstGP$	0.052***	0.047***	0.044***
	(0.002)	(0.001)	(0.002)
Short-Run			
EC term (ϕ_p)	-0.251***	-0.349***	-0.229***
	(0.008)	(0.010)	(0.007)
1st order lag of ΔRC	-0.168***	-0.143***	-0.142***
	(0.007)	(0.007)	(0.006)
$\Delta zPracticePop$	-0.014	0.002	-0.019
	(0.053)	(0.053)	(0.054)
$\Delta zConsPerPat$	0.000	0.001	-0.010***
	(0.002)	(0.002)	(0.002)
$\Delta zDaysPerEstGP$	0.005***	0.004***	0.002+
	(0.001)	(0.001)	(0.001)
$\Delta zShareEstGP$	0.034***	0.030***	0.041***
	(0.002)	(0.002)	(0.002)
Practice FE	Yes	Yes	Yes
Controls	None	SR Only	SR & LR
Observations	29,643	29,643	29,643
AIC	-3.82	-3.89	-3.91

Standard errors in parentheses.

EC refers to the error-correcting speed of adjustment term,

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

5.4 Estimating workforce side variables using different cutoffs

Tables 5.3 and 5.4 report results from changing the threshold for a full work day to 5 and 15 consultations, respectively, as discussed in Section 4.4.2.3 of Chapter 4. Results are consistent with those reported in Chapter 4.

Table 5.3 PMG estimates of the long- and short-run effects on RC – Uses a cutoff of 5 consultations for a full work day.

	(1)	(2)	(3)
Long-Run			
$zPracticePop$	-0.080***	-0.057***	-0.066***
	(0.010)	(0.009)	(0.010)
$zConsPerPat$	-0.017***	-0.011***	-0.032***
	(0.002)	(0.002)	(0.002)
$zDaysPerEstGP$	0.040***	0.029***	0.040***
	(0.002)	(0.001)	(0.002)
$zShareEstGP$	0.052***	0.047***	0.044***
	(0.002)	(0.001)	(0.002)
Short-Run			
EC term (ϕ_p)	-0.251***	-0.349***	-0.229***
	(0.008)	(0.010)	(0.007)
1st order lag of ΔRC	-0.168***	-0.143***	-0.142***
	(0.007)	(0.007)	(0.006)
$\Delta zPracticePop$	-0.014	0.002	-0.019
	(0.053)	(0.053)	(0.054)
$\Delta zConsPerPat$	0.000	0.001	-0.010***
	(0.002)	(0.002)	(0.002)
$\Delta zDaysPerEstGP$	0.005***	0.004***	0.002+
	(0.001)	(0.001)	(0.001)
$\Delta zShareEstGP$	0.034***	0.030***	0.041***
	(0.002)	(0.002)	(0.002)
Practice FE	Yes	Yes	Yes
Controls	None	SR Only	SR & LR
Observations	29,643	29,643	29,643
AIC	-3.82	-3.89	-3.91

Standard errors in parentheses.

EC refers to the error-correcting speed of adjustment term,

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5.4 PMG estimates of the long- and short-run effects on RC – Uses a cutoff of 15 consultations for a full work day.

	(1)	(2)	(3)
Long-Run			
$zPracticePop$	-0.073***	-0.043***	-0.070***
	(0.009)	(0.008)	(0.009)
$zConsPerPat$	-0.023***	-0.017***	-0.033***
	(0.002)	(0.002)	(0.002)
$zDaysPerEstGP$	0.051***	0.037***	0.051***
	(0.002)	(0.002)	(0.002)
$zShareEstGP$	0.042***	0.034***	0.038***
	(0.002)	(0.001)	(0.002)
Short-Run			
EC term (ϕ_p)	-0.252***	-0.350***	-0.234***
	(0.009)	(0.010)	(0.008)
1st order lag of ΔRC	-0.167***	-0.138***	-0.142***
	(0.007)	(0.007)	(0.006)
$\Delta zPracticePop$	-0.015	-0.007	-0.014
	(0.051)	(0.050)	(0.054)
$\Delta zConsPerPat$	0.000	0.001	0.006**
	(0.002)	(0.002)	(0.002)
$\Delta zDaysPerEstGP$	0.006***	0.006***	0.004**
	(0.001)	(0.001)	(0.001)
$\Delta zShareEstGP$	0.032***	0.028***	0.034***
	(0.002)	(0.002)	(0.002)
Practice FE	Yes	Yes	Yes
Controls	None	SR Only	SR & LR
Observations	29,643	29,643	29,643
AIC	-3.81	-3.87	-3.88

Standard errors in parentheses.

EC refers to the error-correcting speed of adjustment term.

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

5.5 Calculating the dependent variable using an alternative tie-breaking method

As described in Section 4.4.2.1 of Chapter 4, we define a patient's regular GP at consultation j as the GP that patient i saw more frequently across all face-to-face consultations with GPs over a two-year time window prior to time t. In Chapter 4, in case of a tie, if that tie includes one or more established GPs, then we randomly select one of those established GPs, else we randomly select one of the unestablished GPs in the tie.

As noted in Section 4.6.4 of Chapter 4, an alternative way of breaking ties is to simply select the GP with whom the patient had their most recent appointment (prior to consultation j) as the regular GP. In order to ensure that the tie-breaking method in the main analysis does not affect our results, we have performed this alternative analysis. Results are given in Table 5.5 and are consistent with the main analysis.

Table 5.5 PMG estimates of the long- and short-run effects on RC – Uses an alternative approach to identify the regular GP in case of ties.

	(1)	(2)	(3)
Long-Run			
$zPracticePop$	-0.074***	-0.053***	-0.073***
	(0.009)	(0.009)	(0.009)
$zConsPerPat$	-0.021***	-0.017***	-0.035***
	(0.002)	(0.002)	(0.002)
$zDaysPerEstGP$	0.051***	0.039***	0.049***
	(0.002)	(0.001)	(0.002)
$zShareEstGP$	0.045***	0.035***	0.038***
	(0.002)	(0.001)	(0.002)
Short-Run			
EC term (ϕ_p)	-0.259***	-0.361***	-0.240***
	(0.008)	(0.010)	(0.008)
1st order lag of ΔRC	-0.162***	-0.134***	-0.136***
	(0.007)	(0.007)	(0.006)
$\Delta zPracticePop$	-0.016	-0.012	-0.027
	(0.057)	(0.057)	(0.059)
$\Delta zConsPerPat$	-0.002	-0.000	0.007**
	(0.002)	(0.002)	(0.002)
$\Delta zDaysPerEstGP$	0.006***	0.005***	0.003**
	(0.001)	(0.001)	(0.001)
$\Delta zShareEstGP$	0.032***	0.028***	0.035***
	(0.002)	(0.002)	(0.002)
Practice FE	Yes	Yes	Yes
Controls	None	SR Only	SR & LR
Observations	29,643	29,643	29,643
AIC	-3.81	-3.88	-3.89

Standard errors in parentheses.

EC refers to the error-correcting speed of adjustment term.

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

5.6 Calculating different lag structures of the PMG model

In columns (1)–(4) of Table 5.6, we report PMG model estimations using different lag structures from those presented in Chapter 4, as discussed in Section 4.6.4. Specifically, from left to right the

columns correspond to PMG$(1,1)$, PMG$(2,2)$, PMG$(3,3)$ and PMG$(4,4)$ models, respectively.

Note first that the AIC values cannot be directly compared across the models, since they do not include the same number of observations (due to differences in the lag structures). Second, observe that the results are consistent with those in the paper.

Table 5.6 PMG estimates of the long-run effects on RC – Using different lag structures.

	(1)	(2)	(3)	(4)
Long-Run				
$zPracticePop$	-0.049***	-0.060***	-0.070***	-0.147***
	(0.007)	(0.009)	(0.010)	(0.012)
$zConsPerPat$	-0.027***	-0.034***	-0.039***	-0.035*
	(0.002)	(0.002)	(0.003)	(0.003)
$zDaysPerEstGP$	0.035***	0.043***	0.051***	0.049***
	(0.002)	(0.002)	(0.003)	(0.003)
$zShareEstGP$	0.033***	0.034***	0.038***	0.044***
	(0.002)	(0.002)	(0.002)	(0.002)
EC term (ϕ_p)	-0.269***	-0.202***	-0.178***	-0.160***
	(0.008)	(0.007)	(0.007)	(0.007)
1st order lag of ΔRC		-0.224***	-0.271***	-0.285***
		(0.008)	(0.011)	(0.012)
2nd order lag of ΔRC			-0.090***	-0.121***
			(0.008)	(0.010)
3rd order lag of ΔRC				-0.037**
				(0.011)
Practice FE	Yes	Yes	Yes	Yes
Controls	SR & LR	SR & LR	SR & LR	SR & LR
Observations	29,963	29,643	29,323	29,003
AIC	-3.93	-3.90	-3.82	-3.79

Standard errors in parentheses.

EC refers to the error-correcting speed of adjustment term.

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

5.7 Alternative model estimation using a micro-panel – GMM

Since we use a methodology that is relatively novel to the OM community, we complement this study by using an alternative modeling strategy to confirm our results.

We organize the data into an unbalanced panel with two levels, practice and time (i.e., year). This yields a micro or short panel (small T, large N) compared to the macro panel (large T, small N) we used earlier. To build the model step by step and to provide rationale for our modeling approach, we start by estimating a time and entity FE regression model specified by

the equation:

$$\begin{aligned} RC_{py} &= \alpha_p + \gamma_y + \beta_1 DaysPerEstGP_{py} + \beta_2 ShareEstGP_{py} \\ &+ \beta_3 ShareGP_{py} + \beta_4 SharePartner_{py} + \boldsymbol{\beta}_5^{\mathsf{T}} \mathbf{X}_{py} + \epsilon_{py} \,. \end{aligned} \tag{5.2}$$

Practice-specific intercepts, α_p, capture unobserved time-invariant heterogeneity across practices. Common time effects, γ_y, capture shocks and trends in RC which affect all practices in the sample. Meanwhile, the vector $\mathbf{X}_{py}$ contains the set of control variables described in Section 5.8.2, and $\epsilon_{py} \sim \mathcal{N}(0, \sigma^2)$ is the idiosyncratic error term. Standard errors are clustered at the practice level to account for autocorrelation within the same practice.

However, as outlined in Section 5.1, the FE (and RE) estimator assumes strict exogeneity, or, equivalently, current values of ϵ_{py} cannot be correlated with past, present, or future values of the predictor variables. The strict exogeneity condition is violated in the presence of correlated time-varying omitted variables, feedback or reverse causality, and dynamic effects. If the serially uncorrelated disturbance assumption is violated in a static regression, then serial correlation has consequences similar to heteroskedasticity (which can be addressed, for instance, by estimating robust standard errors). However, evidence of serial correlation may also be a sign of misspecification of the underlying model, e.g., if the true model is dynamic but is wrongly assumed to be static (Balestra 1982). A model is said to be dynamic if history matters, i.e., if the dependent variable is influenced not only by the current value of the independent variable(s), but also by values of the independent variable(s) in the past. We argue that RC is "sticky" and persistent over time. If these dynamics are present but not sufficiently captured, coefficient estimates may be biased (Campos and Kinoshita 2008). Testing for serial correlation in our FE/FD models using a procedure proposed by Wooldridge (2002) provides evidence to suggest that, indeed, the within-group error terms are serially correlated. One approach to (at least partially) address this issue and that of omitted variable bias more generally is to include the lag of the dependent variable on the right-hand side of the regression as follows:

$$\begin{aligned} RC_{py} &= \alpha_p + \gamma_y + \phi RC_{py-1} + \beta_1 DaysPerEstGP_{py} + \beta_2 ShareEstGP_{py} \\ &+ \beta_3 ShareGP_{py} + \beta_4 SharePartner_{py} + \boldsymbol{\beta}_5^{\mathsf{T}} \mathbf{X}_{py} + \epsilon_{py} \,. \end{aligned} \tag{5.3}$$

In short (i.e., small T) panels such as our setting, introducing the lag on the DV on the RHS of a fixed effects model can lead to a bias in the estimate of ϕ of the order $1/T$ as $N \to \infty$, which is referred to as Nickell's bias (Nickell 1981). This arises as the FE estimator takes the mean of the RC and subtracts it from equation 5.3, giving $RC^*_{py-1} = RC_{py-1} - \overline{RC}_p$, which would be correlated with error term $\epsilon^*_{py-1} = \epsilon_{py-1} - \bar{\epsilon}_p$. The RC_{py-1} term in RC^*_{py-1} will still be correlated with the ϵ_{py-1} term in ϵ^*_{py-1}, violating strict exogeneity, and specifically, this correlation biases the estimates downward.

Alternatively, we can estimate the same model using Ordinary Least Squares (OLS) without the fixed effects i.e., estimate a pooled model. But since the lagged dependent variable and the error term $(\alpha_p + \epsilon_{py})$ will be positively correlated, this will bias the estimate of ϕ upwards. According

to Bond (2002) the fixed effects with the lagged dependent variable (referred to as Dynamic FE) gives us a lower bound for ϕ whereas the dynamic OLS model gives an upper bound for ϕ.

A further concern is the potential for endogeneity bias when using the FE estimator. This can arise if there exists unobserved time-varying factors within a practice that are correlated with both the independent variables and the error term, violating the exogeneity assumption. For example, a change in practice management may affect the degree of prioritization of access over RC, and it might also affect staffing decisions, e.g., the extent to which a practice relies on part-timers or locums. One commonly used approach to address endogeneity concerns in panels is to lag the regressors as follows:

$$\begin{aligned} RC_{py} &= \alpha_p + \gamma_y + \beta_1 DaysPerEstGP_{py-1} + \beta_2 ShareEstGP_{py-1} \\ &+ \beta_3 ShareGP_{py-1} + \beta_4 SharePartner_{py-1} + \boldsymbol{\beta}_5^{\mathsf{T}} \mathbf{X}_{py-1} + \epsilon_{py}\,. \end{aligned} \tag{5.4}$$

However, this replaces the strict exogeneity assumption in FE with the assumption that (i) unobserved variables are serially uncorrelated, and (ii) there are dynamics in the independent variables but not in the dependent variable (Bellemare et al. 2017).

We first report the results from the four estimators described above in Table 5.7 and then proceed to our final modelling strategy that address both endogeneity concerns as well the dynamic panel bias that arises in the previous models.

5.7.1 Results from preliminary models for micropanels

Table 5.7 Preliminary models for micropanels

	FE	FE - lagged regressors	Dynamic Pooled	Dynamic FE
$zPracPop$	-0.037*	-0.026	-0.007***	-0.028*
	(0.018)	(0.021)	(0.001)	(0.013)
$zConsReg$	-0.014*	0.012+	-0.005**	-0.014**
	(0.006)	(0.006)	(0.002)	(0.005)
$zDaysPerEstGP$	0.026***	0.015***	0.011***	0.021***
	(0.004)	(0.004)	(0.002)	(0.005)
$zShareEstGP$	0.040***	0.018***	0.013***	0.033***
	(0.004)	(0.004)	(0.001)	(0.004)
RC_{py-1}			0.842***	0.342***
			(0.010)	(0.038)
Practice FE	Yes	Yes	Yes	No
Controls	Yes	Yes	Yes	Yes
Observations	2,689	2,308	2,308	2,308
R^2	0.508	0.376	0.884	0.545

Notes: $^{+}\ p < 0.10$, $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$; Standard errors, clustered by practice, in parentheses;

Results from the FE regression, the FE regression with lagged regressors, the dynamic pooled OLS model and the dynamic fixed effects model, are reported in Table 5.7. The top panel in the results table reports the coefficients of the variables of interest and the lagged dependent variable. The bottom panel reports the structure of the controls that are included, with "Yes" or "No" indicating inclusion or non-inclusion, respectively. Additionally, we report the R^2 of the model.

Due to the limitations of these approaches, we postpone interpretation of the results for later. However, we note that the direction of the effects of the four independent variables on the RC as well as the statistical significance is consistent across all models (except for the workload variables from the FE - lagged regressor model). Though, by inspecting the coefficients it appears that workforce fragmentation factors are important in explaining the variation in RC relative to the workload factors. Moreover, the coefficients on the lagged dependent variable are large and significant, indicating that history of RC is relevant and important.

5.7.2 Generalized Method of Moments

To address the endogeneity issue as well as the unobserved heterogeneity, we consider the Generalized method of moments by Arellano and Bond (1991), augmented by Arellano and Bover (1995), Blundell and Bond (1998), and Windmeijer (2005). (We closely follow Chung (2017) and Pennetier et al. (2019) to describe the methods used in this section). The idea behind the GMM estimator is to construct instruments from within the panel data structure. More specifically, the method is similar to using lagged differences and levels as instruments as proposed by Anderson and Hsiao (1981, 1982) and does not rely on the use of strictly exogenous instruments. Moreover, this estimator is especially efficient for a small T, large N setting and heteroskedasticity that is prevalent.

Both the difference GMM by Arellano and Bond (1991) and SGMM by Blundell and Bond (1998) and further by Windmeijer (2005) address the limitations mentioned above. In the DGMM estimator, the starting point is to take the first difference of equation 5.3 as follows:

$$\begin{aligned} RC_{py} - RC_{py-1} &= \gamma_y - \gamma_{y-1} + \phi(RC_{py-1} - RC_{py-2}) + \beta_1(DaysPerEstGP_{py} - DaysPerEstGP_{py-1}) \\ &+ \beta_2(ShareEstGP_{py} - ShareEstGP_{py-1}) \\ &+ \beta_3(ShareGP_{py} - ShareGP_{py-1}) + \beta_4(SharePartner_{py} - SharePartner_{py-1}) \\ &+ \boldsymbol{\beta}_5^{\mathsf{T}}(\mathbf{X}_{py} - \mathbf{X}_{py-1}) + \epsilon_{py} - \epsilon_{py-1}\,. \end{aligned} \tag{5.5}$$

or, equivalently,

$$\begin{aligned} \Delta RC_{py} &= \Delta\gamma_y + \phi\Delta RC_{py-1} + \beta_1\Delta DaysPerEstGP_{py} \\ &+ \beta_2\Delta ShareEstGP_{py} \\ &+ \beta_3\Delta ShareGP_{py} + \beta_4\Delta SharePartner_{py} \\ &+ \boldsymbol{\beta}_5^{\mathsf{T}}\Delta\mathbf{X}_{py} + \Delta\epsilon_{py}\,. \end{aligned} \tag{5.6}$$

where $\Delta RC_{py} = RC_{py} - RC_{py-1}$, $\Delta\gamma_y = \gamma_y - \gamma_{y-1}$, $\Delta DaysPerEstGP_{py} = DaysPerEstGP_{py} - DaysPerEstGP_{py-1}$, $\Delta ShareEstGP_{py} = ShareEstGP_{py} - ShareEstGP_{py-1}$, $\Delta ShareGP_{py} = ShareGP_{py} - ShareGP_{py-1}$, $\Delta SharePartner_{py} = SharePartner_{py} - SharePartner_{py-1}$, $\Delta\mathbf{X}_{py} = \mathbf{X}_{py} - \mathbf{X}_{py-1}$ and $\Delta\epsilon_{py} = \epsilon_{py} - \epsilon_{py-1}$.

But, since $RC_{py-1} - RC_{py-2}$ and $\epsilon_{py} - \epsilon_{py-1}$ are correlated because RC_{py-1} and ϵ_{py-1} are correlated, we cannot use a simple OLS model for estimation. Instead, following Anderson and Hsiao (1981, 1982), we can use the lagged differences ΔRC_{py-2} or the lagged level RC_{py-2} as instruments for ΔRC_{py-1} as they are both correlated to $(RC_{py-1} - RC_{py-2})$ but not with the error term, though using lagged levels instead of differences enables the use of a larger number of observations. According to Holtz-Eakin et al. (1988) and Arellano and Bond (1991), all past levels can be used as instrument for ΔRC_{py-1}.For example, RC_{p1} is a valid instrument

for ΔRC_{p2}, RC_{p1} and RC_{p2} can be used for ΔRC_{p3}. Following this pattern, each time period will have a different number of instruments with later periods having more moment conditions available. The matrix of instruments for the lagged continuity variable for practice i would be:

$$Z_p = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ RC_{p1} & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & RC_{p1} & RC_{p2} & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & RC_{p1} & RC_{p2} & RC_{p3} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

But, there are shortcomings of the DGMM method. Using this instrument matrix, there is a loss in degrees of freedom, since the first observation available is at $y = 3$. Moreover, lagged levels may be poor instruments for first differenced variables, especially if variables are close to a random walk.

The solution by Arellano and Bover (1995) and Blundell and Bond (1998) is to create a stacked dataset, where the differences are instrumented by levels and the levels are instrumented with differences. This utilizes more moment conditions and hence more information from the data.

We use the `xtabond2` function in STATA to estimate the two-step SGMM estimator, perform the Hansen (1982) J test for over-identifying restrictions (to test overall exogeneity), the difference-in-Hansen tests (to test the exogeneity of each endogenous variable), the Arellano-Bond test (to identify the start point of each instrument lag structure), and validate that we choose the lowest number of instruments (Roodman 2009a,b).

5.7.3 Results from GMM estimation

Results from the GMM estimation corresponding to the lag of the dependent variable and four main regressors are reported in Table 5.8. The top panel of Table 5.8 reports the short run coefficients, whereas the bottom panels describes the structure of the controls and validity tests.

Column (2) corresponds to the results from the difference GMM estimator whereas Column (3) represents the results from the two-step system GMM estimator.

According to Bond et al. (2001), a consistent estimate of ϕ should be expected to lie between the dynamic pooled OLS estimate ($\phi = 0.388$) and the dynamic fixed effects estimate ($\phi = 0.627$). Specifically, the dynamic FE estimate provides a lower bound for ϕ and the OLS estimate an upper bound. If the DGMM estimate is close to (or below) the FE estimate, this signals that the GMM estimate is biased downward, maybe due to weak instruments, and in this case, SGMM is preferred. Since this holds in our case ($\phi_{DGMM} = 0.388$ $\phi_{SGMM} = 0.627$), we proceed to interpret the coefficients from the SGMM model (Column 3 of Table 5.8)[1].

[1]Consistent results when (1) restricting instruments to first 2 lags only (165 instruments) (2) when using PCA to reduce the instrument count (130 instruments)

Table 5.8 GMM models

Variables	DGMM	SGMM	LR coefficients
1-year lagged dependent variable	0.411***	0.628***	
zPracPop	-0.027	-0.021***	-0.056***
zConsReg	-0.032**	-0.019***	-0.051***
zDaysPerEstGP	0.016	0.015**	0.040**
zShareEstGP	0.018*	0.019***	0.052***
Practice fixed effects	Yes	Yes	
Year fixed effects	Yes	Yes	
Number of observations	1935	2308	
Number of practices	359	372	
Number of instruments	197	285	
Arellano-Bond test for AR(1)[a]	0.000	0.000	
Arellano-Bond test for AR(1)[b]	0.066	0.771	
Hansen J-test for overidentifying restrictions[c]	0.329	0.403	

Notes: [a] $H0$ =no 1st-order serial correlation in residuals

[b] $H0$ =no 2nd-order serial correlation in residuals

[c] $H0$ =model specification is correct

$^{+}\, p < 0.10$, $^{*}\, p < 0.05$, $^{**}\, p < 0.01$, $^{***}\, p < 0.001$;

Column (4) reports the long run coefficients corresponding to the model calculated in Column (3). These coefficients are calculated by scaling the short-run coefficients by the term $(1-\phi)^{-1}$. For example, the long-run effect of $zDaysPerEstGP$ is given by[2]:

$$zDaysPerEstGP_{LR} = \frac{\beta_1}{(1-\phi)}$$

Starting with the workload factors and all else remaining equal, a 1σ increase in $PracPop$ or $ConsReg$ leads to a 5.6 p.p. or 5.1 p.p. reduction in RC, respectively. As for the workforce factors, we find that a 1σ decrease in $DaysPerEstGP$ or $ShareEstGP$ leads to a 4.0 p.p. or 5.2 p.p. reduction in RC, respectively.

To improve interpretation of the results, in Table 5.9 we use the between-practice standard deviation (BPSD) from Table 5.12 to examine how the average RC rate is expected to vary across practices based on practice characteristics.

For example, comparing a large practice with a small practice, where the former has a $PracticePop$ two BPSDs above the mean and the latter two BPSDs below the mean, shows that the larger practice will have a RC rate 40.7% (= (0.326−0.550)/0.550) lower than that of the smaller practice, with RC reduced by 22.4 p.p. Meanwhile, a practice that relies more on part-timers (with $DaysPerEstGP$ two BPSDs below the mean) will only be able to match patients with their regular provider 38.2% of the time. This is 22.7% (49.4-38.2/49.4) lower than a practice that uses predominantly full-time salaried workers, which matches patients with their preferred provider 49.4% of the time. A practice that relies more on unestablished GPs (with $ShareEstGP$ two BPSDs below the mean) will only be able to match patients with their regular provider 35.4% of the time. This is 32% lower than a practice that uses predominantly established workforce, which matches patients with their preferred provider 52.2% of the time.

Table 5.9 Average variation in relational continuity across practices, by practice characteristics.

Variable	-2σ	-1σ	$+0\sigma$	$+1\sigma$	$+2\sigma$
Smaller vs. larger practice population	0.550	0.494	0.438	0.382	0.326
Lower vs. higher consultation rate per patient	0.524	0.481	0.438	0.395	0.352
Fewer vs. more days worked per month by est. GPs	0.382	0.410	0.438	0.466	0.494
Lower vs. higher dependence on est. GPs	0.354	0.396	0.438	0.480	0.522

Notes: This comparison is based on the variation that exists between practices rather than total variation. For example, the $+1\sigma$ column shows the impact of a one unit increase in the between variation for the relevant variable, so for $zConsPerPat$ this would equal the estimated coefficient from Table 4.3, −0.043, multiplied by the between variation reported in Table 5.12, i.e., 0.84.

[2]Note that $(1-\phi)^{-1}$ is the formula for the sum of a geometric series.

5.7.4 Explaining the Trend in Relational Continuity

Next, we investigate which factors are most important in explaining the decline in RC over time. In Table 5.10, for each of the variables we report the change in the value over the observation period. The change is calculated by first estimating the sum of the change in RC over the study period (Column 2), the sum of the change in the independent variables over the study period (Column 3,6,9,12), the sum of the short-run effect of the variable on RC (for each time period the short-run effect of the variable on RC is calculated by multiplying coefficient correspond to the independent variable (from Table 5.8 column 3) times the change in the independent variable for that period). Ultimately, the total long-run effect of the independent variable is calculated by estimating the long-run effect of the independent variable for each period (which is found by calculating the sum of the geometric series up to that period) and summing across all periods.

Table 5.10 Trend in key variables.

Variable		*PracPop*			*ConsReg*			*DaysPerEstGP*			*ShareEstGP*		
Year	Δ %RC	Δ	SR Effect	Total Effect	Δ	SR Effect	Total Effect	Δ	SR Effect	Total Effect	Δ	SR Effect	Total Effect
2008	–	–	–	–	–	–	–	–	–	–	–	–	–
2009	-1.80	0.02	-0.12	-0.32	0.07	-0.38	-1.0	0.04	0.17	0.46	-0.06	-0.31	-0.82
2010	-1.10	0.01	-0.07	-0.18	-0.05	0.26	0.68	-0.08	-0.33	-0.87	-0.04	-0.23	-0.60
2011	-1.30	0.00	-0.02	-0.06	-0.01	0.03	0.07	-0.06	-0.22	-0.58	-0.06	-0.29	-0.75
2012	-1.70	0.01	-0.04	-0.09	0.08	-0.41	-1.03	-0.03	-0.11	-0.28	-0.06	-0.30	-0.76
2013	-0.90	0.02	-0.12	-0.30	-0.05	0.25	0.61	-0.01	-0.04	-0.10	0.04	0.18	0.45
2014	-1.30	-0.01	0.07	0.17	-0.19	0.96	2.17	-0.06	-0.24	-0.54	0.00	0.00	0.01
2015	-1.30	0.02	-0.14	-0.28	-0.11	0.56	1.14	-0.04	-0.15	-0.30	0.00	-0.02	-0.04
2016	-0.40	0.01	-0.03	-0.05	0.04	-0.20	-0.33	0.08	0.32	0.52	0.00	0.00	0.00
2017	-0.80	0.05	-0.26	-0.26	-0.05	0.25	0.25	-0.06	-0.25	-0.25	-0.17	-0.86	-0.86
Σ	-10.60%	0.13	-0.73%	-1.38%	-0.26	1.32%	2.57%	-0.21	0.85%	-1.94%	-0.35	-1.82%	-3.37%
Total reduction explained:		1.38/10.60=13.0%			2.57/10.6=24.2%			1.94/10.6=18.3%			3.37/10.6=31.8%		

Notes: [a] For each variable, we estimate the Δ change in the value from one year to the next, short-run effect of the variable on the %RC for each time period, which is calculated by multiplying coefficient correspond to the independent variable (from Table 5.8 column 3) times the change in the independent variable for that period, and the the long-run effect of the independent variable is calculated by estimating the long-run effect of the independent variable for each period (which is found by calculating the sum of the geometric series up to that period).

Finally, the sum of the total effect for each variable (Column 5,8,11,14) divided by the sum of the total change in %RC (10.6%) allows us to identify the contribution of each of the factors to the total reduction in RC. This shows, for example, that a shift to a more fragmented part-time workforce can explain 18.3% of the 0.106 p.p reduction in RC, whereas an increase in reliance on the non-established workforce can explain 31.8% of the decline in RC. Hence, together, these factors alone can thus explain more than half of the total reduction in RC over the ten-year observation period.

5.8 Alternative analysis - Workforce fragmentation only

For a practice manager, since workforce fragmentation is more easily addressable than workload, I conduct an alternative analysis which only looks at workforce composition as independent variables.

5.8.1 Independent Variables

5.8.1.1 Part-time work by established doctors.

To capture the degree of part-time work at the practice, we first calculate the number of days worked per established doctor in a given week, since it is more common for doctors to work weekly rotas. Then, for each practice-week we average the work days of the established doctors who worked for at least half a day during that week in order to avoid capturing coding errors and special cases, and then finally average over all the full weeks in the given practice-year. A decrease in this measure indicates a shift from full-time to a part-time workforce.

5.8.1.2 Dependence on unestablished doctors.

To calculate the dependence of the practice on the unestablished doctor workforce, we look at the consultations performed by established doctors as a proportion of the total number of consultations performed by any doctor, at practice p in each year y. A decrease in this measure indicates a shift in activity away from the established workforce.

5.8.1.3 Consultation provision by partners versus non-partners.

To calculate the trend away from partners, we calculate the number of consultations provided by partners as a proportion of the number of consultations provided by any established doctor, at practice p in each year y. A decrease in this measure indicates a reduced role of partners, with this activity instead provided by salaried doctors.

5.8.1.4 Task-shifting from doctors to non-doctors.

To capture the degree of task-shifting, we calculate the number of consultations provided by doctors as a proportion of consultations provided by any clinical (doctors and non-doctors) at practice p in year y. A decrease in this measure indicates a greater degree of task-shifting to non-doctors as providers of clinical care.

5.8.2 Control Variables

Various other factors might affect a practice's ability to provide COC and may confound the relationship between COC and the workload and workforce composition variables. The inclusion of practice fixed effects (FEs) in the model accounts for time-invariant factors that are specific to the practice: for example, whether it serves a rural or urban population, population socioeconomic status, etc. Time FEs, meanwhile, can adjust for any factors that change over time and have a common effect on all practices. However, FEs are unable to account for time-varying factors that differ across practices over time. Therefore, we have also defined a number of additional control variables at the practice-year level. The controls are summarized in Table 2.3 and described fully in Appendix Y. Broadly speaking, these controls capture heterogeneity within practices over time in supply (e.g., workforce size), demand (e.g., number of patients), capacity utilization, service times, population demographics, and quality-related outcomes.

Table 5.11 Table of controls.

Variable	Type	Description
Supply and demand		
Labor supply (of clinical staff)	Continuous	The number of full-time equivalent (FTE) clinical staff, calculated weekly and averaged over the practice-year
Patients per clinical staff	Continuous	The number of patients registered at a practice per FTE
Service intensity	Continuous	The annual number of consultations with clinical staff per patient registered at a practice
Capacity utilization	Continuous	The number of consultations divided by the estimated number of consultations that could have been performed, calculated weekly and averaged over the practice-year
Service times		
Appointment duration	Continuous	Average consultation duration*
CV in appointment duration	Continuous	Coefficient of variation of consultation duration*
Population demographics		
Age	Continuous	Average age of patients*
Gender	Continuous	Average percentage of females*
Comorbidities	Continuous	Average number of comorbidities, calculated using the Cambridge Comorbidity Score (Payne et al. 2020)*
Quality-related outcomes		
ED admissions	Continuous	Annual number of emergency admissions by patients on the practice list, averaged over the practice-year
Outpatient referrals	Continuous	The percentage of consultations resulting in an outpatient referral*
New prescriptions	Continuous	The percentage of consultations resulting in a new prescription*

Notes: When a * is included at the end of the description, it indicates that the associated variable is calculated by taking the average across all appointments that occurred within the associated practice-year.

5.8.3 Summary Statistics

Panel A of Table 5.12 contains summary statistics for each of the main variables described in Sections 2.4.3.1 and 5.8.1. To improve interpretation of the results later, we standardize the independent variables and controls by taking their z-scores (i.e., subtracting the mean and dividing by the standard deviation). This is a linear transformation and so has no impact on the results, but coefficients in our models must now be interpreted as the impact on COC of a one standard deviation change in the corresponding variable.

Summary statistics for the standardized variables are reported in Panel B of Table 5.12, followed by a table of correlations in Panel C. The correlation table shows that (except for $zPctDoc$), there is a moderate to strong degree of correlation between the independent variables and COC.

Table 5.12 Descriptive Statistics and Correlations for Variables

Panel A: Descriptive Statistics

					St. Dev.	
	Mean	Median	Min	Max	Overall	Between	Within
COC	0.46	0.43	0.07	1.00	0.17	0.16	0.07
$WorkDays$	3.00	3.00	0.50	5.73	0.58	0.56	0.26
$PctEst$	0.82	0.85	0.10	1.00	0.15	0.14	0.08
$PctPartner$	0.9	1.00	0.00	1.00	0.19	0.19	0.09
$PctDoc$	0.69	0.69	0.03	1.00	0.18	0.18	0.08

Panel B: Descriptive Statistics - Standardized Variables

					St. Dev.	
	Mean	Median	Min	Max	Overall	Between	Within
$zWorkDays$	0.00	-0.03	-4.33	4.74	1.00	0.97	0.45
$zPctEst$	0.00	0.22	-4.69	1.16	1.00	0.90	0.51
$zPctPartner$	0.00	0.54	-4.99	0.56	1.00	1.00	0.46
$zPctDoc$	0.00	-0.01	-5.75	2.74	1.00	0.94	0.48

Panel C: Correlations

	(1)	(2)	(3)	(4)	(5)
(1) COC	1.00				
(2) $zWorkDays$	0.49***	1.00			
(3) $zPctEst$	0.41***	0.31***	1.00		
(4) $zPctPartner$	0.29***	0.22***	0.05**	1.00	
(5) $zPctDoc$	0.07***	0.21***	0.06**	0.03	1.00

$^{+}p < 0.10$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$; St. Dev. short for standard deviation.

This is especially the case for $zWorkDays$, for which the correlation with COC takes value 0.50 ($\mathrm{p} < 0.001$). Meanwhile, the degree of correlation between the independent variables are small to moderate in size and give us no cause for concern, with the variance inflation factors (VIFs) all taking values less than 1.11.

5.8.4 Results

Results from the FE, FE-DPD, and Pooled-DPD estimations are reported in Table 5.13. The top panel in the results table reports the coefficients of the variables of interest and, where appropriate, the lagged dependent variable. The bottom panel reports the structure of the controls that are included, with "Yes" or "No" indicating inclusion or non-inclusion, respectively.

Importantly, we find evidence that the FE estimator may not be appropriate in this setting: The coefficients of the lagged dependent variable are large and significant for both the FE-DPD and Pooled-DPD models, indicating that history is relevant and important in explaining

Table 5.13 Preliminary models

	FE	FE-DPD	Pooled-DPD
$zWorkDays$	0.032***	0.023***	0.011***
	(0.005)	(0.005)	(0.001)
$zPctEst$	0.038***	0.028***	0.012***
	(0.004)	(0.004)	(0.001)
$zPctPartner$	0.021***	0.017***	0.004***
	(0.006)	(0.004)	(0.001)
$zPctDoc$	-0.019***	-0.017***	-0.005***
	(0.006)	(0.005)	(0.001)
COC_{py-1}		0.388***	0.836***
		(0.031)	(0.010)
Practice FE	Yes	Yes	No
Time FE	Yes	Yes	Yes
Controls	Yes	Yes	Yes
Observations	2,832	2,454	2,454
R^2	0.438	0.540	0.884

Clustered standard errors in parentheses.

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

current realizations of COC.[3] However, the two approaches result in very different estimates of this coefficient (0.39 versus 0.84), suggesting potential bias. Moreover, the R^2 of the dynamic models are significantly higher than the static FE model (0.438 in the static FE model versus 0.540 and 0.884 in the dynamic models). Due to the limitations with the three approaches, we therefore postpone interpretation of the results presented in Table 5.13. However, we note that the direction of the effects of the four independent variables as well as the statistical significance is consistent across all models. Furthermore the results are directionally consistent with expectation.

As the FE, FE-DPD and Pooled-DPD estimators are unable to reliably estimate the size of the dynamic effect, and with other sources of endogeneity not yet addressed, we require a new identification strategy. This is the focus of Section 5.7.2.

Table 5.14 System GMM models

	(1)	(2)	(3)	(4)
Short-run coefficients				
COC_{py-1}	0.684***	0.649***	0.619***	0.638***
	(0.086)	(0.103)	(0.085)	(0.072)
$zWorkDays$	0.021***	0.026**	0.023*	0.019*
	(0.006)	(0.010)	(0.009)	(0.009)
$zPctEst$	0.022***	0.015*	0.017*	0.018*
	(0.004)	(0.007)	(0.007)	(0.007)
$zPctDoc$	-0.007**	-0.026**	-0.024*	-0.025**
	(0.003)	(0.009)	(0.010)	(0.009)
$zPctPartner$	0.007**	0.006	0.007	0.006
	(0.002)	(0.005)	(0.005)	(0.005)
Long-run coefficients				
$zWorkDays$	0.068***	0.075***	0.060**	0.053*
	(0.009)	(0.023)	(0.020)	(0.023)
$zPctEst$	0.069***	0.043**	0.044**	0.049**
	(0.013)	(0.016)	(0.016)	(0.016)
$zPctDoc$	-0.022**	-0.074*	-0.063*	-0.070*
	(0.009)	(0.030)	(0.030)	(0.029)
$zPctPartner$	0.023***	0.017	0.018	0.017
	(0.006)	(0.013)	(0.013)	(0.013)
Number of instruments	30	42	54	78
Arellano-Bond test for AR(1)[a] (p-value)	0.000	0.000	0.000	0.000
Arellano-Bond test for AR(2)[a] (p-value)	0.523	0.500	0.521	0.475
Hansen J-test for over-identification[c] (p-value)	0.063	0.076	0.105	0.430
Difference-in-Hansen test of exogeneity[d] (p-value)	0.878	0.185	0.373	0.929

Note:$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Variables considered endogenous:

(1) COC_{py-1}, (2) COC_{py-1}, $zWorkDays$, $zPctEst$, $zPctDoc$, $zPctPartner$

(3) COC_{py-1}, $zWorkDays$, $zPctEst$, $zPctDoc$, $zPctPartner$, Supply and Demand controls (Refer to Table 2.3)

(4) COC_{py-1}, $zWorkDays$, $zPctEst$, $zPctDoc$, $zPctPartner$, all controls (Refer to Table 2.3)

[a] $H0$ = no 1st-order serial correlation in residuals, [b] $H0$ = no 2nd-order serial correlation in residuals

[c] $H0$ = model specification is correct, [d] $H0$ = instruments are exogenous.

5.8.5 GMM Results

Table 5.15 Trend in key variables.

Variable		*WorkDays*			*PctEst*			*PctDoc*		
Year	Δ %COC	Δ	SR Effect	Total Effect	Δ	SR Effect	Total Effect	Δ	SR Effect	Total Effect
2007	–	–	–	–	–	–	–	–	–	
2008	-0.871	-0.042	-0.089	-0.244	-0.025	-0.050	-0.136	0.005	-0.016	-0.044
2009	-1.709	0.065	0.137	0.373	-0.050	-0.100	-0.274	0.013	-0.040	-0.109
2010	-1.205	-0.105	-0.220	-0.593	-0.044	-0.089	-0.240	0.036	-0.108	-0.293
2011	-1.503	-0.106	-0.222	-0.590	-0.040	-0.080	-0.212	-0.007	0.022	0.059
2012	-1.788	0.004	0.008	0.020	-0.063	-0.126	-0.327	0.003	-0.010	-0.027
2013	-0.890	-0.017	-0.037	-0.091	0.028	0.055	0.137	-0.015	0.045	0.112
2014	-1.594	-0.023	-0.048	-0.110	-0.035	-0.070	-0.162	-0.001	0.003	0.007
2015	-1.275	-0.065	-0.137	-0.281	-0.081	-0.161	-0.330	-0.071	0.214	0.438
2016	0.007	-0.038	-0.079	-0.130	-0.045	-0.091	-0.149	0.052	-0.155	-0.255
2017	-2.039	-0.137	-0.287	-0.287	-0.121	-0.242	-0.242	0.053	-0.160	-0.160
Σ	-12.866%	-0.422	-0.886	-1.933	-0.452	-0.904	-1.936	0.063	-0.190	-0.271
Total reduction explained:		1.933/12.866=15.026%			1.936/12.866=15.044%			0.271/12.866=2.105%		

Notes: [a] For each variable, we estimate the Δ change in the value from one year to the next by running a time and entity fixed effects model of the form: $Y_{py} = \alpha_p + \mu_y \lambda_y + \epsilon_{py}$. The coefficient of the time fixed effect for each of the time periods μ_y captures the yearly COC for that period relative to the first time period, using which we calculate the Δ change in each time period. Meanwhile, α_p is a practice FE that controls for the fact that the starting values of the dependent variables may differ across practices and also for the fact that some practices drop out of the sample during the observation period, influencing μ_y. The short-run effect of the variable on the %COC for each time period is calculated by multiplying coefficient correspond to the independent variable (from Table 5.14 column 4) times the change in the independent variable for that period, and the the long-run effect of the independent variable is calculated by estimating the long-run effect of the independent variable for each period (which is found by calculating the sum of the geometric series up to that period).

Finally, the sum of the total effect for each variable (Column 5,8,11) divided by the sum of the total change in %COC (12.87%) allows us to identify the contribution of each of the factors to the total reduction in COC. This shows that a shift to a more fragmented part-time workforce can explain 15.026% of the 0.128 p.p reduction in COC, whereas a shift to the non-established workforce, can explain 15.04% of the reduction in COC A physician-dominant workforce can explain a further 2.1% of the COC reduction. Together, these three factors alone can thus explain nearly 32.18% of the total reduction in COC over the eleven-year observation period.

[3] In the Pooled-DPD model, we find that past values of COC beyond the first lag is insignificant, hence we proceed by including only one lag of COC in further models.

References

Anderson TW, Hsiao C (1981) Estimation of dynamic models with error components. *Journal of the American statistical Association* 76(375):598–606.

Anderson TW, Hsiao C (1982) Formulation and estimation of dynamic models using panel data. *Journal of econometrics* 18(1):47–82.

Arellano M, Bond S (1991) Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies* 58(2):277–297.

Arellano M, Bover O (1995) Another look at the instrumental variable estimation of error-components models. *Journal of econometrics* 68(1):29–51.

Balestra P (1982) Dynamic Misspecification and Serial Correlation. *Qualitative and Quantitative Mathematical Economics*, 115–145 (Springer, Dordrecht).

Baltagi B (2008) *Econometric analysis of panel data* (John Wiley & Sons).

Baltagi B, Griffin JM (1984) Short and long run effects in pooled models. *International Economic Review* 25(3):631–45.

Bellemare MF, Masaki T, Pepinsky TB (2017) Lagged explanatory variables and the estimation of causal effect. *The Journal of Politics* 79(3):949–963.

Blundell R, Bond S (1998) Initial conditions and moment restrictions in dynamic panel data models. *Journal of econometrics* 87(1):115–143.

Bond SR (2002) Dynamic panel data models: a guide to micro data methods and practice. *Portuguese economic journal* 1(2):141–162.

Bond SR, Hoeffler A, Temple JR (2001) Gmm estimation of empirical growth models. *Available at SSRN 290522* .

Campos N, Kinoshita Y (2008) Foreign direct investment and structural reforms: Panel evidence from eastern europe and latin america. *IMF Staff Papers* .

Chung DJ (2017) How much is a win worth? an application to intercollegiate athletics. *Management Science* 63(2):548–565.

Gokpinar B, Hopp WJ, Iravani SM (2010) The impact of misalignment of organizational structure and product architecture on quality in complex product development. *Management science* 56(3):468–484.

Hansen LP (1982) Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society* 1029–1054.

Holtz-Eakin D, Newey W, Rosen HS (1988) Estimating vector autoregressions with panel data. *Econometrica: Journal of the econometric society* 1371–1395.

Nickell S (1981) Biases in dynamic models with fixed effects. *Econometrica: Journal of the Econometric Society* 1417–1426.

Payne RA, Mendonca SC, Elliott MN, Saunders CL, Edwards DA, Marshall M, Roland M (2020) Development and validation of the cambridge multimorbidity score. *CMAJ* 192(5):E107–E114.

Pennetier C, Girotra K, Mihm J (2019) R&d spending: Dynamic or persistent? *Manufacturing & Service Operations Management* 21(3):636–657.

Reed WR, Zhu M (2017) On estimating long-run effects in models with lagged dependent variables. *Economic Modelling* 64:302–311, ISSN 02649993.

Roodman D (2009a) How to do xtabond2: An introduction to difference and system gmm in stata. *The stata journal* 9(1):86–136.

Roodman D (2009b) A note on the theme of too many instruments. *Oxford Bulletin of Economics and statistics* 71(1):135–158.

Samargandi N, Fidrmuc J, Ghosh S (2015) Is the relationship between financial development and economic growth monotonic? evidence from a sample of middle-income countries. *World development* 68:66–81.

Smith J, Holder H, Edwards N, Maybin J, Parker H, Rosen R, Walsh N (2013) Securing the future of general practice New models of primary care Summary. Technical report, Nuffield Trust, URL `www.nuffieldtrust.org.uk/publications/securing-future-general-practice`.

Ul Haque N, Pesaran MHH, Sharma S (2005) Neglected Heterogeneity and Dynamics in Cross-Country Savings Regressions. *SSRN Electronic Journal* URL `http://dx.doi.org/10.2139/ssrn.267794`.

Windmeijer F (2005) A finite sample correction for the variance of linear efficient two-step gmm estimators. *Journal of econometrics* 126(1):25–51.

Wooldridge JM (2002) Econometric analysis of cross section and panel data mit press. *Cambridge, MA* 108.

Chapter 6

Conclusions

This thesis has made contributions to a growing body of research that is concerned with the repeated interactions between a patient and a doctor and how this can be used as an operational lever to improve the efficiency and effectiveness of primary care service provision.

The first study in this thesis titled "Continuity of care increases clinical productivity in primary care" furthers our understanding of the implications of repeated patient-doctor interactions and the long-term relationship between them on direct productivity effects in primary care practices themselves. We find that appointments with a regular family doctor who offers continuity of care as opposed to appointments with an unknown "transactional provider" affect the productivity of primary care practices. This is an essential question as continuity of care is eroding in the wake of demand outpacing the supply of doctors. This trend towards transactional consultations is further exacerbated by the emergence of online primary care providers with massive scale. Offering patients appointments with their regular doctor rather than the next available clinician will be costly for the practice, unless the regular doctor provides a service that reduces sufficient future demand from that patient to outweigh the additional cost. We demonstrate that care continuity has substantial productivity benefits in primary care because patients who see their regular doctor will visit less often. Revisit intervals following consultations with regular doctors are substantially longer, extending the time between consultations by approximately 13% after accounting for selection effects. Therefore, primary care practices that seek to increase productivity by increasing daily throughput (e.g., by pooling demand and scheduling ever more consultations into a doctor's day) are in fact reducing productivity if this is done at the cost of care continuity, as these patients will start returning more often. We also demonstrate that the productivity benefit of continuity of care is strongest for older patients and patients with complex chronic conditions. The findings suggests that the future of primary care services may be a natural segmentation of the service into:

a. A transactional service for those relatively healthier patients for whom continuity of care makes little difference and for whom daily throughput is the appropriate productivity metric, and;

b. A relational service (epitomized by the notion of the "family doctor") for those patients with complex chronic needs that need one trusted doctor where "the buck stops". This approach prevents these patients from "bouncing from one doctor to the next" without ever getting "properly sorted out" (quotes from doctors in primary care practices).

The second study in this thesis, titled "The Operational Determinants of Relational Continuity

of Care" helps practice managers to identify the root cause of low rates of continuity at their own practice, the key operational levers that they can use to promote continuity, and proposes strategies to mitigate the adverse effects of industry trends on continuity. It also provides valuable insights for policymakers and health service commissioners who wish to incentivize primary care practices to safeguard continuity. We find that a sustained increase in workload caused by growth in demand relative to supply and increasing fragmentation of the workforce due to a shift to part-time and agency work can explain nearly 50% of the decline in continuity over the past decade. These factors also cause significant heterogeneity to exist between practices in their ability to provide continuity. Therefore, when workload and workforce changes are unavoidable, GP practice managers should be aware of the potential adverse effects on continuity provision and adopt proactive strategies to minimize these effects. Diminishing returns from pooling also suggest that a creative middle ground between dedicated queues and full pooling could be explored. One example of this would be a situation in which two part-time GPs working offset shifts emulate one full-time GP by sharing the responsibility for one set of patients. In this case, continuity is established not with a single provider but with a defined pair of providers. In the nursing context, this dual-provider setup has been successfully used and is sometimes referred to as the "pod" model (Friese et al. 2014). Future research might explore the extent to which this approach can shelter patients against the adverse affects associated with a loss of the continuity established between a patient and a single provider. Our results can help explain why some practices are less able to provide continuity than others. In particular, we observe, based on both our own interactions with practice managers and on statements from professional bodies representing primary care providers (Jeffers and Baker 2016), that continuity is widely recognized as a cornerstone of the delivery model used in general practice. However, providers are often unaware of the root causes of low rates of continuity within their own practices. Our paper can help practice managers to answer this question by characterizing the impact of practice characteristics on continuity (see, e.g., Table 4.4), thus enabling them to design targeted interventions to improve continuity provision. We also note that the decline in continuity is an issue that patients care about: the majority of patients value interpersonal care continuity, yet many report that they are unable to see their preferred provider (Aboulghate et al. 2012). Comparing the operational characteristics of practices can thus help patients, especially those who value continuity most highly, to choose the practice that is best for them.

We note also that although this work is focused on the primary care context, it provides valuable insights and methodologies that may also prove useful in the study of other knowledge-intensive or relationship-based services (e.g., banking, legal, consultancy, etc.). In such settings, knowledge accumulation and learning that occurs through repeated interactions can help to improve service outcomes, quality, and efficiency (Henderson and Tookes 2012, Taylor and Plambeck 2007, Huckman and Pisano 2006). However, exploiting such benefits requires firms to be able to match customers with servers on an ad-hoc basis as and when demand materializes. We demonstrate the critical role that staff scheduling and the use of in-house (rather than external)

service providers can play in facilitating these repeated interactions. Moreover, we provide an estimation approach that can be adopted in other contexts for identifying the relative importance of different factors in explaining a firm's ability to provide customer-server continuity.

With respect to methodology, we introduce to the OM community the panel ARDL modeling framework which, until now, has primary been used by researchers in the field of macroeconomics. As highlighted in Section 4.6, ARDL models have many advantages over traditional panel methods such as the FE and FD estimators. In particular, in a dynamic setting panel ARDL methods enable direct estimation of the long-term or permanent impact of the independent variables on the LR equilibrium while also addressing endogeneity bias through inclusion of firm-specific fixed effects and lags of the dependent and independent variables in the SR estimation equation. We anticipate that this novel approach can be easily extended to other empirical OM contexts, both within and outside of the healthcare domain, especially as researchers are increasingly gaining access to large macro panels (i.e., large N and large T).

Overall, as the first paper to demonstrate empirically the importance of operational factors in driving variation in continuity between practices and over time, this work provides important insights for practice and provokes a range of follow-up questions that might be pursued in future research.

As future implications of this work, I wish to look further in to what can primary care practices can do to make their physicians more productive, in other words, deliver the same or better quality of care with less physician time. Specifically, with increasing demand and shrinking workforce, I wish to look at what the tipping point is at which the patient panel becomes too large and physician's become less productive? Once this is established, the next question would be, what managerial levers would we have to shift this tipping point? What type of organizational support and process helps clinicians look after more patients? (e.g. locums to take on transactional consultations, better triaging, less non-clinical administration, others)

Theoretically, an increased patients-to-GP ratio reduces the supply per patient and therefore extends the frequency of visits as patients are more often unable to access the GP (the system fills up more often). This leads to poor patient experience for all as patients but will not lead to poorer outcomes for all as some patients will get better with time and will not need to see their GP. However, it will lead to poorer outcomes for some, who will then (a) turn up in ED, either as "worried well" (discharged) or because they have an urgent clinical need (admitted to hospital), and (b) may require more GP visits later.

From an operations perspective, there are several possible responses to an increasing constraint of a critical resource (GP time)

- Prevention: Manage demand pro-actively without using GP time, so that patients don't need to request appointments

- Triage: Help receptionists to channel appointment requests appropriately, so that the most value-add patients get access to the increasingly constrained resource (GP time)
- Specialization: E.g. having GPs who will do only urgent care / transactional medicine on the day (they may also be involved in triaging)
- Shift some of the tasks that GPs perform to other staff (nurses, emergency care practitioners, administration) to allow the GP to "work at the top of her license"

Another potential research avenue would be to ask: What do practices do differently to keep healthy patients healthy for longer? Specifically, I would identify a point in each of the patients' pathways when they become 'unhealthy' (develop one of the 38 comorbidities) and investigate the differences between the patient journey of such 'unhealthy' patients and 'healthy' patients. I would expect that one such variable would be the frequency of visits or the engagement of the patient with the practice and whether this level of engagement influences the probability of the patient being diagnosed with one of the chronic diseases. Moreover, it will be both clinically and operationally meaningful to know whether the point that the patients become unhealthy can be delayed by increasing the level of engagement with the practice.

References

Aboulghate A, Abel G, Elliott MN, Parker RA, Campbell J, Lyratzopoulos G, Roland M (2012) Do english patients want continuity of care, and do they receive it? *British Journal of General Practice* 62(601):e567–e575.

Friese CR, Grunawalt JC, Bhullar S, Bihlmeyer K, Chang R, Wood W (2014) Pod nursing on a medical/surgical unit: Implementation and outcomes evaluation. *Journal of Nursing Administration* 44(4):207–211, ISSN 15390721, URL `http://dx.doi.org/10.1097/NNA.0000000000000051`.

Henderson BJ, Tookes H (2012) Do investment banks' relationships with investors impact pricing? the case of convertible bond issues. *Management Science* 58(12):2272–2291.

Huckman RS, Pisano GP (2006) The firm specificity of individual performance: Evidence from cardiac surgery. *Management Science* 52(4):473–488.

Jeffers H, Baker M (2016) Continuity of care: still important in modern-day general practice. *British Journal of General Practice* 66(649):396–397.

Taylor TA, Plambeck EL (2007) Supply chain relationships and contracts: The impact of repeated interaction on capacity investment and procurement. *Management science* 53(10):1577–1593.