INSEAD
The Business School for the World®

Working Paper Series
2020/47/TOM

Explaining the Erosion of Relational Care Continuity: An Empirical Analysis of Primary Care in England

Harshita Kajaria-Montag
University of Cambridge, hk437@cam.ac.uk

Michael Freeman
INSEAD, michael.freeman@insead.edu

Relational continuity (RC) refers to an ongoing relationship between a patient and a clinician or clinical team beyond a specific service encounter or disease episode. As a defining characteristic of good medical practice, RC has been shown to confer many clinical and operational advantages and is desired by patients, clinicians, and policymakers alike. Yet despite its many benefits, RC in the primary care setting has been in sharp decline over the past decades, contributing to poorer health outcomes and lower efficiency as well as falling patient and provider satisfaction. Anecdotally, this downward trend has been attributed to a sustained increase in workload caused by a growing and aging population and to changes in the workforce composition caused by burnout and workload pressures. However, there is a dearth of evidence to support or contradict this impression, and the key factors that cause changes in RC are not well understood. As a result, little action is being taken to slow or reverse this trend. We fill this gap by empirically examining the main operational factors that can explain variation in RC both between practices and over time. To do so, we use a unique dataset of primary care consultations corresponding to '10% of England's population over ten years. Using a panel ARDL estimation approach, we show that workload and workforce factors have a significant influence on a primary care practice's ability to provide RC, explaining '36% of the residual within-practice variation after inclusion of all other controls. We also find that three factors alone can explain '45% of the decline in RC over the study period: increasing fragmentation of the workforce caused by i) primary care practitioners shifting to part-time work patterns and ii) greater dependence on temporary staff; a sustained increase in workload caused by iii) greater patient volumes without a proportionate increase in physician-hours. Of these, workforce factors appear to be relatively more important than workload factors, with increasing workforce fragmentation driving '33% of the total decline. We discuss the implications of these findings for patients and primary care practice managers, and we suggest strategies for maintaining levels of RC in the face of these industry trends.

Key words : Healthcare; Relational Continuity; Primary Care; Econometrics; Panel ARDL
History : September 25, 2020

Electronic copy available at: http://ssrn.com/abstract=3699385

Explaining the Erosion of Relational Care Continuity: An Empirical Analysis of Primary Care in England

Harshita Kajaria-Montag
Judge Business School, University of Cambridge, Cambridge CB2 1AG, United Kingdom hk437@cam.ac.uk
Michael Freeman
INSEAD, 1 Ayer Rajah Avenue, Singapore 138676, michael.freeman@insead.edu,

Relational continuity (RC) refers to an ongoing relationship between a patient and a clinician or clinical team beyond a specific service encounter or disease episode. As a defining characteristic of good medical practice, RC has been shown to confer many clinical and operational advantages and is desired by patients, clinicians, and policymakers alike. Yet despite its many benefits, RC in the primary care setting has been in sharp decline over the past decades, contributing to poorer health outcomes and lower efficiency as well as falling patient and provider satisfaction. Anecdotally, this downward trend has been attributed to a sustained increase in workload caused by a growing and aging population and to changes in the workforce composition caused by burnout and workload pressures. However, there is a dearth of evidence to support or contradict this impression, and the key factors that cause changes in RC are not well understood. As a result, little action is being taken to slow or reverse this trend. We fill this gap by empirically examining the main operational factors that can explain variation in RC both between practices and over time. To do so, we use a unique dataset of primary care consultations corresponding to $\simeq$10% of England's population over ten years. Using a panel ARDL estimation approach, we show that workload and workforce factors have a significant influence on a primary care practice's ability to provide RC, explaining $\simeq$36% of the residual within-practice variation after inclusion of all other controls. We also find that three factors alone can explain $\simeq$45% of the decline in RC over the study period: increasing fragmentation of the workforce caused by i) primary care practitioners shifting to part-time work patterns and ii) greater dependence on temporary staff; a sustained increase in workload caused by iii) greater patient volumes without a proportionate increase in physician-hours. Of these, workforce factors appear to be relatively more important than workload factors, with increasing workforce fragmentation driving $\simeq$33% of the total decline. We discuss the implications of these findings for patients and primary care practice managers, and we suggest strategies for maintaining levels of RC in the face of these industry trends.

Key words: Healthcare; Relational Continuity; Primary Care; Econometrics; Panel ARDL
History: September 25, 2020

1. Introduction

Relational continuity (RC) refers to "a therapeutic relationship between a patient and one or more providers [...] that spans various healthcare events and results in accumulated knowledge of the patient and care consistent with the patient's needs" (Burge et al. 2011). In contrast to the episode-focused secondary care model in which "diseases stay and patients come and go," RC is a service concept especially important in the primary care setting where, by contrast, "patients

stay and diseases come and go" (Heath 1995). In particular, for the patient, having access to a familiar provider who they know and trust and can turn to when they are feeling unwell can lead to better health outcomes and improved experience (Dossa et al. 2017). Meanwhile, for the clinician, repeated interactions can increase the 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, Freeman et al. 2010). RC is thus advocated as a cornerstone of primary care and an essential element of general practice.

Despite the many reported benefits of greater RC, however, the trend over the past decade has moved in the opposite direction. For example, patient-perceived RC was found to have declined by 27.5% for primary care patients in the UK between 2012 and 2017 (Levene et al. 2018), with similar trends observed in the US (Fletcher et al. 2011, Ladapo and Chokshi 2014). Several reasons have been put forward to explain this decrease in RC in primary care settings, which broadly can be attributed to (i) a sustained increase in workload and (ii) dependence on an increasingly fragmented workforce, factors which are both exerting intensifying pressure on primary care practices. These same factors are also thought to be critical in explaining the growing variation *between* practices in their ability to provide RC.

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. By 2032, for example, there is projected to be a shortage of 21 to 55 thousand primary care physicians (PCPs) in the US (Heiser 2019). Meanwhile, in England, the number of general practitioners (GPs, the UK equivalent of a PCP) has already fallen from a high of 66 per hundred thousand population in 2009 to just 58 in 2018, with this downward trend projected to continue (Palmer 2019). These intensifying workload pressures have also had a knock-on effect on the workforce composition. Many GPs have sought to improve their working conditions and counter burnout by going part-time or working more flexibly as locums (i.e., stand-in doctors who are non-salaried and paid per shift that they work). In the face of workforce shortages, practices have also found it harder to hire and retain staff, so they have increasingly turned to temporary workers in order to fill gaps. This has resulted in a more fragmented workforce comprising fewer full-time salaried doctors catering to greater numbers of more complex patients. Patients thus find it more difficult to make an appointment with their regular provider, contributing to the decline in RC.

While these explanations for the loss of RC may be intuitive, it should be emphasized that, to the best of our knowledge, they are only anecdotal, and no empirical evidence exists to corroborate

them. Moreover, in the wake of these pressures, individual practices now vary substantially in their ability to provide RC, and it is still unclear which factors are most important in explaining this variation. Yet identifying the causes of trends and variation in RC is becoming increasingly important against the backdrop of a growing interest in continuity of care (COC) within the healthcare and operations management communities.[1] For example and as discussed in more detail in Section 2, recent studies by Senot (2019), Ahuja et al. (2020b), and others have begun exploring the relationship between care continuity and patient outcomes. These studies collectively suggest that health managers and policymakers should try to provide greater continuity for specific patient populations. However, while these findings are helpful, it remains challenging to devise effective strategies and policy recommendations without first acquiring a fundamental understanding of the key levers that can be used to promote continuity. Meanwhile, for patients who prioritize care continuity, there is little guidance on which features are important when selecting a new practice.

This paper sets out to fill this gap in the literature by empirically investigating the extent to which workload and workforce composition factors contribute to variation in RC. 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.

[1] Note that the terms RC and COC are sometimes used interchangeably. However, the American Academy of Family Physicians defines COC more broadly as "the process by which the patient and his/her physician-led care team are cooperatively involved in ongoing health care management toward the shared goal of high-quality, cost-effective medical care"(American Academy of Family Physicians 2015). As such, the term COC is typically assumed to encompass three different forms of continuity: informational, management, and relational (Haggerty et al. 2003). In this paper we focus on the latter, while noting that most of the studies that we cite that use the term COC also only consider the RC component.

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 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 $\simeq$61% of the time, more than double the $\simeq$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 $\simeq$48% in January 2008 to $\simeq$36% in December 2017, a relative decrease of $\simeq$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 $\simeq$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 $\simeq$33% to $\simeq$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 7.

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.

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.

The benefits of repeated interactions with the same provider 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).

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.

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.

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.

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 GP 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 GP practice. Coverage is nearly universal, with 98% of the UK population registered with a GP 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 GP 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.

GP 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 GP 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 4.2.3, 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.

3.2. Workload-Related Factors Affecting Relational Continuity

Several factors have contributed to increasing workload pressures faced by GP 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 $\simeq$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

[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.

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). GP 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 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. *Given a fixed number of established GPs, an increase in the number of registered patients will reduce the level of relational continuity.*

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

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 satisfaction

[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.

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 2016).

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. *An increase in the proportion of part-time work within the established workforce will reduce the level of relational continuity.*

HYPOTHESIS 4. *An increase in the unestablished workforce as a proportion of the overall GP workforce will reduce the level of relational continuity.*

4. Data and Variable Descriptions

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

4.1. Data Preparation

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 COC 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.

4.1.2. 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 GP 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 GP 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.

4.2. Variable Descriptions

We next describe the calculation of the key variables included in our study. All summary statistics are provided in Table 1.

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 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.

[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 EC.2 of the e-companion.

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.:

$$RC_{pm} = \frac{\sum_{j \in C_{pm}} RegGP_{(it)j}}{|C_{pm}|}, \tag{1}$$

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

Figure 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.

4.2.2. 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 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, $PracticePop_{pm}$.

[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.

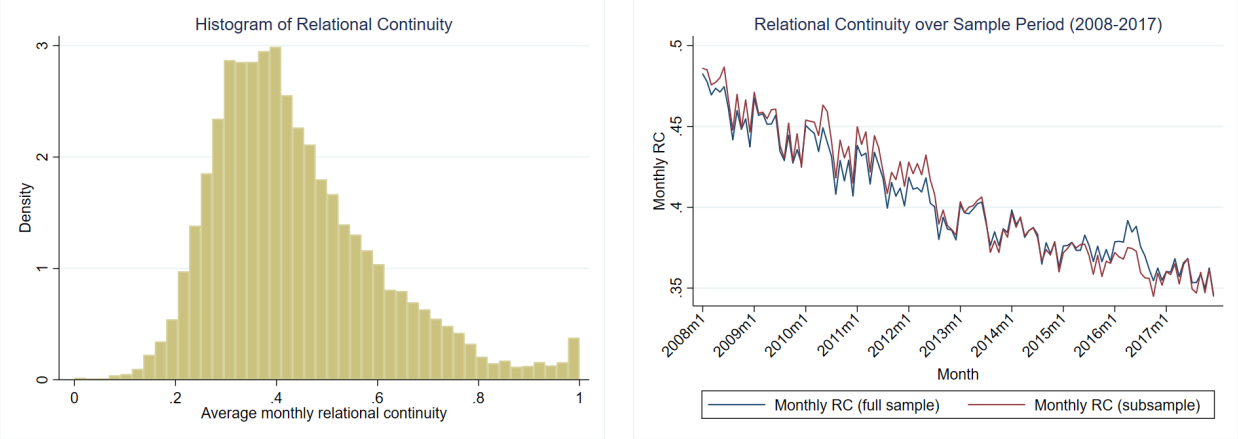

Figure 1 Histogram of relational continuity (left) and trend over time (right), with trend calculated using the weighted average of the full sample (blue line) and the subsample of practices operating during all ten years (red line).

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.

4.2.3. 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 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 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 EC.3 of the e-companion).

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 4.2.2, $GPDaysPerCons_{pm}$ 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 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

[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.

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.

4.3. Summary Statistics

Panel A of Table 1 contains summary statistics for each of the main variables described in Sections 4.2.1 through 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 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 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.

5. Fixed Effects Models

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:

$$RC_{pm} = \alpha_p + \gamma_m + \beta_1 PracticePop_{pm} + \beta_2 ConsPerPat_{pm}$$

Table 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.

$$+ \beta_3 DaysPerEstGP_{pm} + \beta_4 ShareEstGPs_{pm} + \boldsymbol{\beta}_5^{\mathsf{T}} \mathbf{X}_{pm} + \epsilon_{pm} \,. \tag{2}$$

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 4.2.4, 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.

5.2. Results

Results from the FE regression are reported in Table 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

Table 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.

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.

Due to the limitations of the FE approach, which we discuss next in Section 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 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 6.2.

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 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 EC.1 of the e-companion.

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 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.

A further concern is the potential for endogeneity bias when using the FE estimator. This can arise if there exist 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. Our alternative modeling approach helps to resolve these endogeneity concerns.

6. Autoregressive Distributed Lag Models

6.1. Panel ARDL Estimator

To address the limitations described in Section 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

$$RC_{pm} = \alpha_p + \sum_{j=1}^{J} \lambda^*_{pj} RC_{p(m-j)} + \sum_{k=0}^{K} \boldsymbol{\delta}^{*\top}_{pk} \mathbf{Z}_{p(m-k)} + \epsilon_{pm}\,, \tag{3}$$

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):

$$\Delta RC_{pm} = \alpha_p + \phi_p \left[RC_{p(m-1)} - \boldsymbol{\beta}_p^{\mathsf{T}} \mathbf{Z}_{p(m-1)} \right] + \sum_{j=1}^{J-1} \lambda_{pj} \Delta RC_{p(m-j)} + \sum_{k=0}^{K-1} \boldsymbol{\delta}_{pk}^{\mathsf{T}} \Delta \mathbf{Z}_{p(m-k)} + \epsilon_{pm} \qquad (4)$$

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).[7] If ϕ equals zero, then the existence 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 5.3 and described further in Section EC.1 of the e-companion. 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 (4).[8] First is the mean groups (MG) estimator, a fully heterogeneous model that does not impose any

[7] 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).

[8] 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).

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 (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 (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 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.)

6.2. Results

Results from the PMG estimation corresponding to the four main regressors are reported in Table 3. The top panel of Table 3 reports the LR coefficients, whereas the middle panel reports the

Table 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.

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.

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:

$$\Delta RC_{pm}=\alpha_p+\phi_p\left[RC_{p(m-1)}-\boldsymbol{\beta}^{\mathsf{T}}\mathbf{Z}_{p(m-1)}\right]+\lambda_p\Delta RC_{p(m-1)}+\boldsymbol{\delta}_p^{\mathsf{T}}\Delta\mathbf{Z}_{pm}+\epsilon_{pm}\,. \tag{5}$$

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

Table 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 3, -0.037, multiplied by the between variation reported in Table 1, i.e., 0.84.

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 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 we use the between-practice standard deviation (BPSD) from Table 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.

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 5 we report the change in the value of each of the main variables over the observation

Table 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 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.

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 $\left|C^{+}_{pm}\right|$.[9] 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.

Using the above approach to estimate μ, in the Δ column of Table 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 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 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 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%.

[9] 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.

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 4.2.3. 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 EC.2-EC.4 of the e-companion, with all findings consistent.

In addition, while the lag structure in Section 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 EC.5 of the e-companion, indicating that our findings are robust against the presence of potential omitted variables.

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, GP practice managers and 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 2016, 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,[10] 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.

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.

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, RC 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 RC established between a patient and a single provider.

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 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

[10] 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.

impact of practice characteristics on RC (see, e.g., Table 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.

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.

Acknowledgments

This study is based in part on data from the Clinical Practice Research Datalink obtained under licence from the UK Medicines and Healthcare products Regulatory Agency. The data is provided by patients and collected by the NHS as part of their care and support. The interpretation and conclusions contained in this study are those of the author/s alone. This study was approved by the CPRD Independent Scientific Advisory Committee (ISAC) committee. Protocol no.: 19_004R2.

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* 22(5).

American Academy of Family Physicians (2015) Continuity of Care, Definition of. URL `https://www.aafp.org/about/policies/all/continuity-of-care-definition.htmlhttps://www.aafp.org/about/policies/all/definition-care.html`.

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`.

Baltagi B, Griffin JM (1984) Short and long run effects in pooled models. *International Economic Review* 25(3):631–45.

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.

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* 3629032.

Bostock N (2018) More than 1,100 GP practices have closed or merged under NHS England. URL `https://www.gponline.com/1100-gp-practices-closed-merged-nhs-england/article/1497606`.

Bostock N (2019a) Less than three in 10 GPs work full time, official data show. URL `https://www.gponline.com/less-three-10-gps-work-full-time-official-data-show/article/1522455`.

Bostock N (2019b) Number of GP practices in England falls below 7,000. *GPonline* URL `https://www.gponline.com/number-gp-practices-england-falls-below-7000/article/1525443`.

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

Burge F, Haggerty JL, Pineault R, Beaulieu MD, Lévesque JF, Beaulieu C, Santor DA (2011) Relational continuity from the patient perspective: comparison of primary healthcare evaluation instruments. *Healthcare Policy* 7(Spec Issue):124.

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

Centre for Workforce Intelligence (2014) In-depth review of the general practitioner workforce. URL `https://www.gov.uk/government/publications/in-depth-review-of-the-general-practitioner-workforce`.

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. URL https://www.telegraph.co.uk/news/2018/08/16/average-gp-now-works-35-days-week-just-one-20-trainees-plans/.

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.

Fletcher KE, Sharma G, Zhang D, Kuo YF, Goodwin JS (2011) Trends in inpatient continuity of care for a cohort of medicare patients 1996–2006. *Journal of Hospital Medicine* 6(8):438–444.

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

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.

Gault B (2019) Average GP waiting times exceed two weeks for first time ever — News Article — Pulse Today.

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, Okane 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, URL https://www.aamc.org/news-insights/press-releases/new-findings-confirm-predictions-physician-shortagehttps://news.aamc.org/press-releases/article/2019-workforce-projections-update/.

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.

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. Available at `https://www.instituteforgovernment.org.uk/publication/performance-tracker-2019/general-practice` (2020/21/07).

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.

Ladapo J, Chokshi D (2014) Continuity of care for chronic conditions: threats, opportunities, and policy. *Health Affairs Blog* 2016.

Legraien L (2019) Four in 10 patients offered same-day GP appointments. URL `http://www.pulsetoday.co.uk/news/gp-topics/access/four-in-10-patients-offered-same-day-gp-appointments/20039040.article`.

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 — News Article — Pulse Today. URL `http://www.pulsetoday.co.uk/news/gp-topics/employment/gp-practices-locum-use-surges-20-in-a-year/20020002.article`.

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.

NHS England (2016) General practice forward view. Technical report, NHS England.

NHS England (2018) NHS england. Technical report, NHS England, URL `https://www.england.nhs.uk/five-year-forward-view/next-steps-on-the-nhs-five-year-forward-view/primary-care/`.

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. Available at `https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland` (2020/22/07).

Palmer B (2019) Is the number of GPs falling across the UK? URL `https://www.nuffieldtrust.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. Survey, The Physicians Foundation, URL `https://physiciansfoundation.org/wp-content/uploads/2018/09/physicians-survey-results-final-2018.pdf`.

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.

E-companion to:
"Explaining the Erosion of Relational Care Continuity: An Empirical Analysis of Primary Care in the UK"

EC.1. 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 6 of the main paper.

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 EC.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

$$\begin{aligned}\Delta RC_{pm} = {} & \gamma_m + \beta_0 \Delta RC_{p(m-1)} + \beta_1 \Delta PracticePop_{pm} + \beta_2 \Delta ConsPerPat_{pm} + \beta_3 \Delta DaysPerEstGP_{pm} \\ & + \beta_4 \Delta ShareEstGPs_{pm} + \boldsymbol{\beta}_5^{\mathsf{T}} \Delta \mathbf{X}_{pm} + \Delta \epsilon_{pm} \,. \end{aligned} \tag{EC.1}$$

Table EC.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 EC.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 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 (EC.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 3.1. 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 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 (EC.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 5.3 of the main paper.

EC.2. Subsample of 79 practices

Table EC.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 4.1.2 of the main paper. Results are consistent with those reported in the main paper. (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.)

EC.3. Estimating workforce side variables using different cutoffs

Tables EC.3 and EC.4 report results from changing the threshold for a full work day to 5 and 15 consultations, respectively, as discussed in Section 4.2.3 of the main paper. Results are consistent with those reported in the main paper.

EC.4. Calculating the dependent variable using an alternative tie-breaking method

As described in Section 4.2.1 of the main paper, 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 the main paper, 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 6.4 of the main paper, 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 EC.5 and are consistent with the main analysis.

Table EC.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			
$zPracticePop$	-0.062***	-0.063**	-0.079***
	(0.018)	(0.017)	(0.019)
$zConsPerPat$	-0.004	-0.013***	-0.010*
	(0.003)	(0.004)	(0.004)
$zDaysPerEstGP$	0.057***	0.050***	0.059***
	(0.004)	(0.003)	(0.004)
$zShareEstGP$	0.043***	0.031***	0.040***
	(0.003)	(0.002)	(0.003)
Short-Run			
EC term (ϕ_p)	-0.207***	-0.307***	-0.191***
	(0.013)	(0.016)	(0.013)
1st order lag of ΔRC	-0.174***	-0.153***	-0.144***
	(0.013)	(0.012)	(0.012)
$\Delta zPracticePop$	-0.017	0.127	0.025
	(0.086)	(0.088)	(0.080)
$\Delta zConsPerPat$	0.000	0.002	0.009*
	(0.003)	(0.003)	(0.003)
$\Delta zDaysPerEstGP$	0.007***	0.004*	0.003
	(0.002)	(0.002)	(0.002)
$\Delta zShareEstGP$	0.033***	0.030***	0.037***
	(0.003)	(0.003)	(0.003)
Practice FE	Yes	Yes	Yes
Controls	None	SR Only	SR & LR
Observations	9,322	9,322	9,322
AIC	-3.98	-4.04	-4.04

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

EC.5. Calculating different lag structures of the PMG model

In columns (1)–(4) of Table EC.6, we report PMG model estimations using different lag structures from those presented in the main paper, as discussed in Section 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 EC.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 EC.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.
$^{+}$ $p<0.10$, * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

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

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

References for e-Companion

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).

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

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.

Nickell S (1981) Biases in dynamic models with fixed effects. *Econometrica: Journal of the Econometric Society* 1417–1426.

Reed WR, Zhu M (2017) On estimating long-run effects in models with lagged dependent variables. *Economic Modelling* 64:302–311, ISSN 02649993.

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`.

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