The Effect of Reducing Costly Waiting Times on Demand for NHS Services

Anna E.W.

March 2021

The Effect of Reducing Costly Waiting Times of Demand for NHS Services

Executive Summary

Can excessive NHS waiting times be reduced without significantly stimulating demand? Excessive waiting times for healthcare are costly to both the patient's wellbeing and to the healthcare provider in the form of medico-legal costs and reputation damage. England's National Health Service (NHS) healthcare is free at the point of delivery, and as a consequence, long waiting lists have been a feature since it's creation in 1948. We can conceptualise time spent waiting in the queue as a non-money price that patients pay to receive treatment. Basic demand theory suggests that when the 'cost' of NHS treatment falls, consumers are able to 'purchase' more of the good, stimulating levels of demand. An important question is therefore; what happens to the demand curve for NHS treatment throughout the waiting time? Following Martin and Smith (1999) I employ a 2SLS approach to model demand utilisation for NHS elective care. I present estimation results using 5,482 observations for the effect of waiting times on demand for NHS treatment using monthly panel data for 191 Clinical Commissioning Group (CCG) geographical areas between 2017 and 2019.

The results of this study are reassuring. This dissertation reports that the effect of waiting times on demand is negligible, though positive and statistically significant, suggesting that ambitious policy to reduce wait times should be pursued because concerns that demand will be significantly stimulated are unfounded. However, the results also suggest caution, and that there is a complex interaction between supply and demand which requires more complex modelling, but which is beyond the scope of this dissertation.

Supervisors:

Acknowledgements

I thank my supervisor Xiaoshan Chen for her support and guidance, which has been greatly appreciated.

Contents

Declaration

I declare this dissertation is solely the result of my own work. Except where stated otherwise by reference or acknowledgment, the work presented is entirely my own. This dissertation is approximately 11,109 words.

List of Figures

List of Tables

Abbreviations

NHS National Health Service

CCG Clinical Commisioning Group

GP General Practioner

2SLS Two-Stage Least Squares Estimation

OLS Ordinary Least Squares

1 Introduction

1.1 Overview

Since it's creation in 1948, delivering timely healthcare has been a major challenge for the English National Health Service (NHS). A question of high priority is therefore; can waiting times be reduced without significantly stimulating demand? The NHS is the taxpayer-funded government agency provider of healthcare services. It dominates the healthcare market in England where private health care is expensive by international standards. The NHS is also of immense cultural importance and typically talk of privatisation is met with strong disapproval from the public. However, high levels of demand have at times threatened the NHS ability to deliver their own standard of care when supply becomes inadequate. Health inequalities are pervasive in England, and areas of England that are served by under-performing NHS trusts are those with weaker local economies. It should also be noted that the NHS is the largest employer in England, and for this reason alone should be considered as having considerable influence over the economic status of local areas and the country as a whole (Kings Fund, 2020).

Currently, waiting list lengths are at a twelve-year high , growing in all specialities and across all regions. Even before significant Coronovirus-19 hospitalisations there were 4.4 million patients on the waiting list, approximately 730,000 of whom had waited longer than 18 weeks (Morgan, 2021). With several trusts not reporting data, these figures might be higher. They suggest an increasing unmet need in the context of a long term deterioration in performance against key waiting time standards, and ultimately, a breach of commitment to patient care under the NHS Constitution. Costs to public health include the development of co-morbidities, increased mortality and

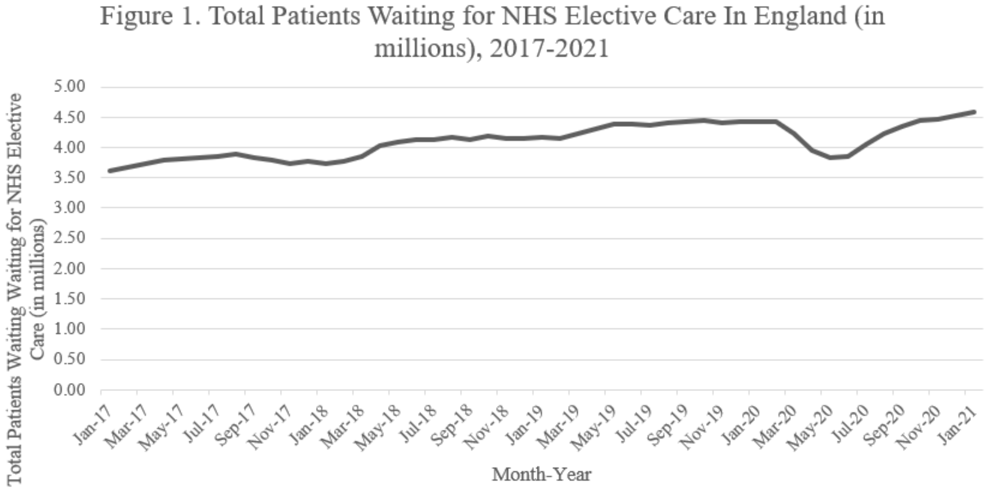

Figure 1: Total Patients Waiting for NHS Elective Care In England (in millions), 2017-2021. Source: NHS, 2021

poorer outcomes that are directly the result of long waiting times. Costs to the NHS are reputational damage and large legal bills for damages that are the direct result of excessive wait times (Kings Fund, 2020), funds for which must come out of healthcare providers annual budgets, which have not expanded in line with the needs of the population. The obvious threat is that of privatisation, which the English public finds unacceptable, and which would inevitably be associated with the access and logistical issues seen in those markets elsewhere in the world.

A substantial injection of funding to reduce waiting times in 2002 later became part of an aggressive effort to reduce wait times from ‘18 months to 18 weeks’. Other system reforms financially rewarded trusts for staying within waiting times targets. These efforts made some headway but have not been sustained. In the period just prior to the influx of Coronovirus-19 hospitalisation, it was estimated that to clear the backlog 500,000 additional patients per year would need to be treated, for 4 years (Kings Fund, 2017).

On one hand, demand theory would appear to encourage us to hypothesise that the problem cannot be solved and that we are doomed to long wait

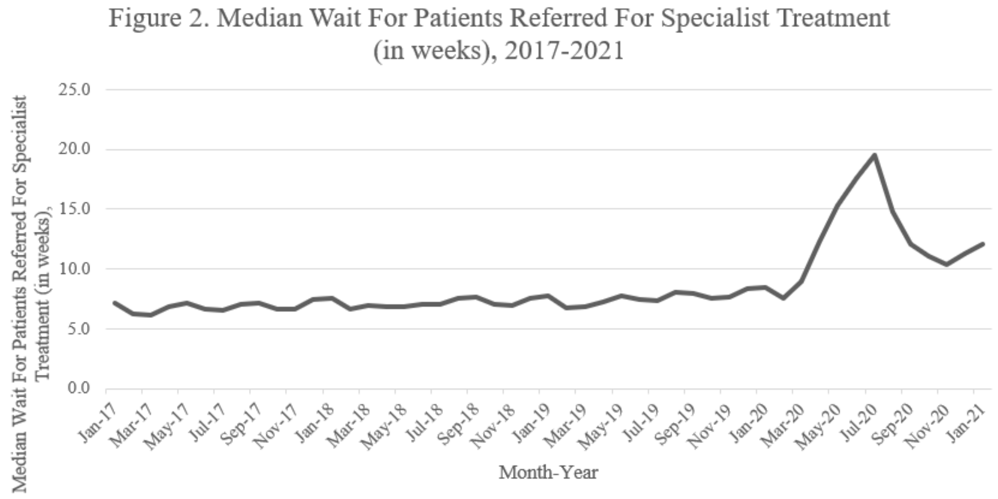

Figure 2: Median Wait For Patients Referred For Specialist Treatment In England (in weeks), 2017-2021. Source: NHS, 2021

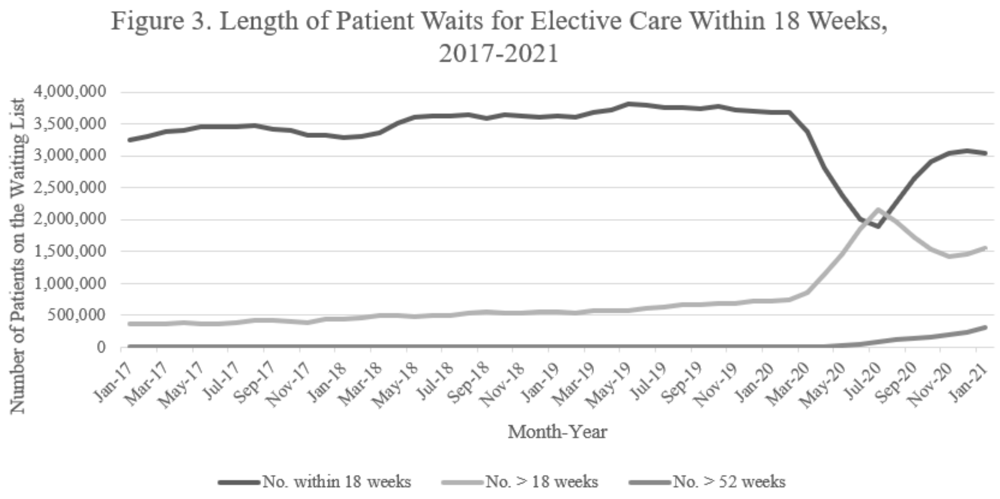

Figure 3: Length of Patient Waits for Elective Care Within 18 Weeks,, 2017-2021. Source: NHS, 2021

times, perhaps an inevitable feature of the Beveridge model. On the other, the mixed results in the literature suggest that the relationship is not so straight-forward.

The benefits of reducing waiting times are obvious and apply not only to patients, but also on a larger economic level if public health is improved. Results similar to Martin and Smith (1999) would suggest that waiting times can be reduced, without NHS resources becoming overwhelmed. A result indicating that reducing waiting times stimulates demand in a significant way would be very problematic for patients, the NHS and policy makers alike.

NHS waiting lists have featured high on the priority list of successive governments. The government plans a large-scale restructuring of the NHS in 2021, and leaked documents (Blackall, 2021) imply the abolition of the commissioner-provider relationship in it's current form. The results of this dissertation could shed light on the efficiency of the current structure of the NHS in the period before a scheduled reorganisation, and could be useful in informing design decisions. The Clinical Commissioning Group (CCG) structure is soon to be dismantled, yet very little independent econometric analysis has been published since their creation in 2012.

1.2 Aims

This dissertation aims to answer the question; can costly NHS waiting times be reduced without significantly stimulating demand? Following Martin and Smith (1999) econometric analysis is applied to monthly panel data for 191 CCGs between 2017 and 2019. This dissertation adds to a very sparse literature on waiting times and demand within the current organisational structure, in fact there are no independent econometric studies on the effect of waiting times on demand within the current structure that the author is aware of. There are currently only a handful of lower-quality studies on waiting time policy effectiveness as a result, and it is clear that our underdeveloped understanding of how waiting times and demand interact are a hindrance to effective policy.

Further to the usual caveats that must be kept in mind when interpreting the results of econometric modelling, the data available for this dissertation are aggregated at the CCG level and this is something I am powerless to change. It is an improvement on some of previous research where data are aggregated at the regional level, but there does exist finer-sampled data that might reveal relationships not seen at a higher level of aggregation.

1.3 Dissertation Outline

This paper is structured as follows; the remainder of this section provides some background on queues as a rationing mechanism. In section 2, the referral process in England is sketched and the organisational structure of the NHS is introduced. In section 3 I review the literature and introduce the theoretical model of NHS demand which is analogous to the standard economic model of demand. Section 4 describes the sample and its source that were available for the study and provides some rudimentary descriptive statistics. Section 5 introduces the dependent, independent and control variables and details their construction. Section 6 introduces the theoretical framework and empirical model of demand used in this study. Section 7 presents and discusses the results of the empirical estimation. I conclude with a reflection on policy implications, possible directions for future research and useful lessons learned in the course of this dissertation.

2 Background On The English National Health Service

2.1 Clinical Commissioning Groups

In 2013 211 Clinical Commissioning Groups (CCGs) replaced Primary Care Trusts and were introduced as part of the Health and Social Care Act 2012. CCGs are geographically bounded entities made up of General Practices (GPs) who commission healthcare for their local population. Annual financial allocations to CCGs are used to purchase healthcare services from providers such as hospitals, community care facilities or private sector providers. CCGs are now responsible for approximately 60 percent of the NHS budget. In this dissertation I analyse waiting times at the CCG area level. During the period of study, there were 191 CCGs in England made up of approximately 8,000 general practices individually commissioning care such as emergency, maternity, elective hospital and community care for approximately 226,000 patients.

2.2 NHS Referrals

The NHS Constitution gives patients the legal right to access treatment for routine, non-urgent procedures within 18 weeks of referral. Patients present their health complaints to their local general practitioner, who then makes the decision whether or not to refer the patient to a NHS consultant specialist at a hospital. As such, GPs are gatekeepers of NHS resources. Once a patient has been referred, they are invariably placed on a waiting list for treatment. Typically, a patient will receive a communication from the hospital informing them of their personal waiting time.

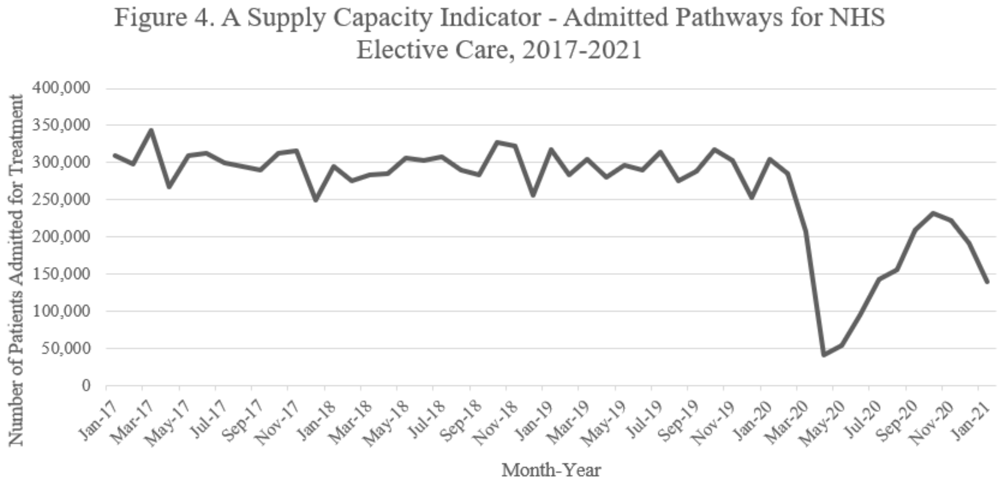

Figure 4: These are patients who have come off the waiting list and have been admitted for treatment, an indicator of the NHS's supply capacity. Source: NHS, 2021

One in twenty GP consultations results in a referral to a specialist. Typically, this happens when a GP feels specialist input is necessary. Clinical responsibility for the patient is transferred from the GP to a consultant. Patients are referred in order to establish a diagnosis, access elective surgery, undergo specialist investigation and for reassurance (Knottnerus, 1991; Daly and Collins, 2007; Fuat et al, 2003; Evans et al, 2007; Bjerager et al, 2006; Dovey et al, 2002). Referral rates vary substantially across the country, with implications for the empirical measurement of demand. An interpretation of these variable rates is difficult to ascertain because their cause is multifactorial. Variables such as population health needs and area deprivation may drive referral rate variation as might the general practitioners individual attitude toward risk and patient pressure. As such, the GP acts as a gatekeeper to specialist care (Kassirer and Kopelman, 1989; Singh et al, 2007; Elstein et al, 2002).

2.3 A Note on Coronavirus-2019

In December 2019 there emerged reports of a pneumonia of unknown microbial aetiology in Wuhan City, Hubei Province, China, with clinical presentations ranging in severity from very mild symptomatology to respiratory failure. Occasionally, severe central nervous system complications have been reported. After the acute stage, a varied picture of long-term sequela of infection was reported in a significant number of patients. An acute infectious respiratory disease caused by a novel coronovirus was found to be the cause. The disease was Coronavirus-2019, also know as SARS-CoV-2, and cases have since been reported across the world. Consequently, a pandemic has formally been declared at the time of writing this dissertation. The pandemic has had a significant effect on the care-seeking behaviour of the public, as well as diverting significant resources away from most areas of healthcare speciality in order to treat a high number of Coronovirus-19 infections. The result has been a historically unprecedented increase in waiting list lengths and wait times. In order to accommodate this, significant ad-hoc organisational change has taken place.

The British Medical Association estimated that between April and December 2020 there were 2.7 million fewer elective procedures, and 18.66 million fewer outpatient attendances (British Medical Association, 2020). This reactivity on the part of the NHS and the public is a response to the temporal patterns of the outbreak. As such, modelling the dynamics of NHS demand as it relates to waiting times during the Coronovirus-19 pandemic is extremely challenging and consequently outside the scope of this dissertation. A study period of 2017-2019 was chosen in order to avoid methodological complications arising from the inevitably numerous structural breaks in the data as a result of variation in the infection rate.

3 Literature Review

The first section of this literature review gives an account of the theoretical background on waiting lists. The second section presents the theoretical framework used in this dissertation. The third section is a review of the empirical literature which attempts to answer the question of whether waiting times affect demand behaviour. Some non-healthcare examples which are worth mentioning are also discussed.

In markets where resources are limited, queues form. Far from representing dysfunction, they serve to reconcile supply and demand in situations where goods have no money-price, as is frequently the case in markets for public goods. In economic theory queues arise as a consequence of non-stochastic variation in markets where price is a below the market clearing level. In markets for public goods this is the result of the imposition of external price constraints. Demanders wait sometimes in physical queues (Dean and Sostiele, 1985), or put their names on a waiting list and queue in absentia (Lindsay and Feigenbaum, 1984) and available supply is rationed. The academic literature is smaller than would be expected for a topic of such importance, and publishing is rather irregular with no recent studies exploring the relationship between demand and waiting times since the introduction of the current NHS structure in April 2013.

In the formal model of waiting times put forward by Lindsay and Feigenbaum (1984), the patients position in the queue is linked to the date on which treatment is commenced and a consequence of this is that value, and therefore demand, is determined by the length of the queue. Because patients join the waiting list costlessly in absentia and are free to simultaneously pursue other opportunities, there is a diminishing value to the good they wait for, as opposed to an increasing cost. The patient looks at the queue length, or the waiting time as it's proxy, and joins the line, or is dismayed by the queue length and refuses to join. Once the patient begins treatment, they are removed from the waiting list.

Queues are an every-day experience in a world of limited resources. You join a queue of traffic to access your email server, as is famously said of London busses- you wait for ages and then three come at once, you call for a doctor's appointment at eight thirty on the dot and an automated messages informs you that you are fifteenth in line. However, queues need not be the result of a 'broken system'. In fact their presence may signal efficiency, and their absence overcapacity. Queuing Theory is the study of waiting in line, and the associated behaviours, gains and losses of individual joiners and leavers. Queues are of special interest to economists and they arise in two generalised settings. The first is as a corollary of a stochastic process, where the number of individuals arriving in the queue in a particular time interval is a Poisson probability distribution. The economist would say that such stochastic events are exogenous to the model. The second scenario in which queues form is of that when a good is priced below the market clearing level, such as public goods that face exogenous price constraints. For example, NHS healthcare, the price of which is not determined by demand, but by a government sponsor. Naturally, economists will ask how rational agents interact with such queues, and how efficiency and fairness can be maximised. A natural development within the literature has been to attempt to assign a cost to the time an individual spends waiting in the queue, of course, if we can do this then we can design optimal systems that impose minimum costs on demanders.

3.1 Theoretical Background

A large body of literature on queuing exists and out of this has grown a substantial application to economics (Cox and Smith, 1961; Cooper, 1981). Queue formation interests economists for it's application to markets where instantaneous price or output adjustment on one or both sides of the market is unfeasibly costly, such as in the setting of oligopoly (De Vany 1976). Another scenario in which queues form is that when a good is priced below the market clearing level, such as public goods that face exogenous price constraints. In this case, queues function as a rationing mechanism. The NHS waiting list is one such queue.

The theoretical framework in this dissertation is taken from Feigenbaum and Lindsay's (1984) work on the rationing mechanism of waiting lists. In their model, V represents a valuation that an individual places on the health gains from medical treatment. If treatment is delayed by time t, then a decay factor operates on V, and is represented by

$$e^{-gt} \tag{1}$$

where g is a decay rate that might reasonably by thought of as being composed of continued pain and suffering, loss of income, development of co-morbidities etc. The value of NHS treatment, received after a wait t, to an individual can therefore be represented as

$$Ve^{-gt} \tag{2}$$

Geographic variations in income, general and acute health need, age distribution and other preferences will in turn cause V and g to vary.

Clearly, the net benefit to an individual is reduced by waiting through the decay function, which can be conceptualised as either a cost to the patient, or as a reduction in the benefit of treatment, perhaps as the result of a deteriorating prognosis. Essentially, the net benefit to the patient receiving treatment is reduced in the presence of a wait t. Additionally, there is an explicit cost C of obtaining treatment (search costs for NHS care or money-costs for private care). Therefore, a patient's utility function can be written as

$$U(V, g, t, C) = Ve^{-gt} - C \tag{3}$$

Within Welfare Economics, there is a large literature on government programs , with central issues of distribution, property rights, fairness for societies poorest and those with the highest needs. Barzel (1974) examines queues from a distributional perspective. Barzel considers a theoretical situation in which a population of individuals has access to a good on which has been imposed price controls such that it is offered at a zero money price via the rationing mechanism of a queue. The cost of waiting is explained as one which arises because no well-defined property rights to the 'free good' exist. Demanders establish such rights by waiting in a queue for however long is necessary - waiting times represent the cost of establishing property rights. Barzel claims that a demand curve can be derived by observing the relationship between quantity demanded and the time-per-unit waiting price. When the per-person quantity demanded is held constant, as the time-price increases, so does the number of demanders. The economic problem that Barzel sketches is analogous to the one faced by the NHS, because the right to receive treatment within 18 months as per the NHS Charter only comes into legal effect when the patient joins the queue for specialist treatment. This result has, in principle, a major implication for waiting time policy design; it seems that waiting times cannot be reduced without increasing the rate at which individuals join the queue.

De Vany (1976) proposes a theoretical model of capacity utilisation queues in the context of oligopoly, argued that rational customers only join queues that have a wait time less than or equal to some critical value. For queues otherwise too long, demanders balk and leave. Hence, suppliers prices and capacity directly influence effective demand for a good, with a higher cost to the individual having a negative effect on the rate of arrivals. Whilst no special attention is paid to the type of good that demanders queue for, this result in the special case of oligopoly is especially relevant given that the NHS dominates the market for healthcare in England.

3.2 Empirical Literature

A smaller body of empirical literature exists on the relationship between waiting times and demand, which is somewhat surprising given the political importance of waiting lists for public healthcare.

Focusing on the demand side of the market, to examine how queues form to ration supply, Lindsay and Feigembaum (1984) applied their theoretical model of queuing to empirical data on mean waiting times and discharges collected from fourteen administrative areas in 1974. They observe waiting lists for each hospital region, which they treat as separate queues. They found that the rate of joining is negatively related to expected delay and to the rate at which demand decays over time. Cullis, John and Jones (1986) respond to Lindsay and Feigenbaum. They argue that welfare costs are difficult to estimate but that costs are not overwhelming, despite waiting list numbers being used as a "big stick" with which to castigate the NHS. One 'common sense' interpretation is that if healthcare services are provided at zero money cost then demand and therefore waiting lists will become infinitely large and that real, finite waiting lists are an artefact of such pricing. The authors disagree that time is a price, rather it is a cost that is imposed on the patient. They highlight evidence that longer waiting lists in fact suggest an increased proportion of patients who do not need treatment (for example, patients who have already been treated privately or the condition has resolved spontaneously), and argue that therefore the Lindsay-Feigenbaum results should be approached with caution. They also draw attention to the fact that Lindsay and Feigenbaum use highly aggregated data at the regional level which does not allow for intra-regional variability in waiting times and queue lengths, after all, waiting lists are managed by hospitals and not by regions. The Lindsay Feigenbaum model has attracted further criticism regrading the claim that patients can ‘shop-around’ for treatment at NHS hospitals. At the time of publication, there was no evidence to support this assumption, rather the contrary. Today however new policy offers patients a choice of waiting lists to join. There is some evidence that the ‘Patient Choice’ policy increases waiting times (Sá et al, 2019).

Cullis, John and Jones use private healthcare procedure costs in 1984 as a proxies for mean waiting times to calculate ball-park welfare costs for patients, and taking a cross section of the fourteen healthcare administrations for 1978 together with waiting time and waiting list data for all diagnostic specialities perform a linear regression analysis. They were unable to replicate Lindsay and Feigenbaum’s results, finding a small positive relationship between mean wait times and demand. They concluded that a finer grained sample is necessary, and that the Lindsay-Feigenbaum results do not have a straightforward interpretation.

Later, Propper (1990), tested the hypothesis as claimed by Lindsay and Feigenbaum that time spent on waiting is costless, the implication being that waiting times should not affect demand for healthcare services. Proper used data for three months in 1987, surveying 1,360 patients. A probit model was employed in an analysis that asked participants at what monetary value they would be willing to trade their place in the queue. It was found that an estimated average value to patients of a reduction of one month in waiting times for non-urgent treatment was £40 and hence, waiting is costly and patients use a cost calculation to determine whether they partake in queuing.

In a study not concerning healthcare, but worth mentioning alongside Propper, Deacon and Sonstelie (1985) conducted a revealed preference experiment and find that rationing by waiting list is socially wasteful, and that it's cost is a deadweight loss. The authors studied motorists fuel purchasing behaviour in a natural experiment of motorists who visited a low-cost petrol station in the state of California in the United States. The motorists chose between waiting in line for low-priced fuel, or purchasing high-priced fuel without waiting. The researchers used their observations to money-price motorists waiting time. The value of waiting time in most cases was close to individuals wages net of tax. They estimate that waiting exhausts 116 percent of the value of the good. Because Propper and Deacon and Sonstelie provide experimental results, where the researchers were able to directly observe queuing phenomena with controls, they lend a special credibility to the theoretical framework that forms the basis of much of the ongoing research.

However, Culyer and Cullis (1976), focusing on supply variables, find no such relationship. They find no evidence that patients were paying a time price for hospital treatment. They conclude that any attempt to regulate waiting times by manipulating supply factors, such as number of beds, clinician headcount, number of specialised operating theatres is unlikely to be effective because they do not find any 'systematic and reliable' relationship between the variables that warrants a plausible behavioural explanation, which is suggestive of an underdeveloped understanding among researchers of the mechanisms at work.

Worthington (1987) presented results of an investigation into the implications of certain waiting list management actions by hospital managers in a linear regression analysis of NHS data collected over 18 months across 9 waiting lists . It was found that increased supply leading to reduced wait times in turn stimulated demand because GPs are encouraged to refer more patients until the waiting time returns to 'normal'. The process is conceptualised as a 'feedback' mechanism. The results support the theory that the arrival rate of patients in to the queue slows as the waiting list size increases, and highlights the fact that incentives and GP behavior must be considered in a model of waiting times.

Also considering incentive structures, Frost (1980) in a long term study between 1949-76 using NHS time-series data for general surgery for which the econometric formulation of the hypothesis (an increase in consultants will cause waiting time to fall but waiting list length to increase because patients are attracted to the lower waiting costs of treatment) distinguished short and long run effects to investigate why waiting lists are longer than we expect. Frost found the size of waiting list to be positively sensitive to the number of hospital consultants, with a 1 percent increase in consultants resulting in a 1 percent increase in the waiting list. Frost concluded that waiting lists arise because consultants are able to control their own work load. It could also be hypothesised that because consultants often have private practices a long waiting list might financially benefit them as exasperated patients leave the NHS queue and seek private treatment. No direct evidence of demand response to lowered waiting times is provided, however alongside the findings of Worthington and Cullis, John and Jones it is clear that future hypothesis will have to extend beyond simple mechanisms of demanders finding lower costs more attractive.

Martin and Smith (1999) propose supply and demand models analogous to standard econometric price based models for waiting lists for elective surgery within the NHS between 1991 and 1992 for 4,460 small geographical areas with England, across twelve clinical specialities using anonymised data from the Hospital Episodes Database. Starting from the theoretical model proposed by Lindsay and Feigenbaum that assumes equilibrium, they employ a two-stage least squares approach and account for temporal and small area effects, to address the issue of waiting time endogeneity. Spatial variation in general and acute need, access to general practitioners, costs of private healthcare and age distribution were modelled. They find low elasticity of demand with respect to waiting times. The implication is that increasing supply in order to reduce waiting times would appear to not greatly stimulate demand. They conclude that the effect is modest, which implies that very significant resources would be necessary to reduce waiting times. Whilst they note difficulties with finding suitable instruments for their analysis, they provide a useful baseline approach to modelling demand that will employed later in this dissertation.

In a very similar exercise, Martin et al (2003) use logistic regression to examine determinants of prolonged wait times for elective treatment, and find factors contributing to waiting lists include throughput, capacity, intensity of activity, population size of the Health Authority, need, and the characteristics of trusts by modelling both the supply and demand side of the market. They found little evidence that NHS supply capacity was related to longer wait times. They do however find increased bed occupancy to be related to prolonged wait times. This result is somewhat counter-intuitive, and highlights the econometric challenges of modelling waiting lists and utilisation, by suggesting a subtle interaction of supply and demand. For example, Culyer and Cullis (1976) theorise that because GPs make hospital referrals, and there is evidence that when the waiting list is long GPs hesitate to refer their patients, then demand and supply become related in a way too complex for simple models to disentangle. Clearly, simple correlations are insufficient and there may be no straightforward relationship between supply and demand variables, and waiting times.

Blundell and Windmeijer (2000), using the same waiting times data for acute care as Martin and Smith (1999) between 1990 and 1991 model the determinants of demand for acute services by employing a probit maximum likelihood estimator and conclude that waiting times are a “hassle-cost”. They remove geographical areas with high wait times from their analysis to avoid systemic measurement error and complications arising from the likely complex interaction of needs variables with demand in those areas with high wait times. Using differences in waiting times across geographies at the small area level, the authors identify the determinants of demand as GP accessibility, hospital accessibility, the proportion of the population of pensioner age and private hospital accessibility, employment, and car ownership.

Gravelle et al (2003) estimate a supply and demand model using Health Authority data on admissions to acute hospitals for 8,414 small geographical areas in England from 1987-1993 for elective surgery and find waiting times highly significant to demand utilisation. Information on waiting times, access, supply capacity, the availability of private healthcare, mortality, education and income were augmented from individual level survey data. A rather discouraging interpretation of these results is that increasing supply would only serve to stimulate demand (Pope, 1992; Roland and Morris, 1988), and policy efforts to reduce waiting lists would exacerbate the situation.

In summary, the theoretical literature frames queuing as a necessary rationing mechanism for public goods in markets analogous to the market for NHS healthcare. Disconcertingly, there is general agreement that if waiting times are reduced, demand will be stimulated. The empirical results are more mixed, indicating an underdeveloped understanding of the complexity of NHS queue behaviour and challenges with modelling demand for healthcare that have yet to be overcome.

3.3 Intended Contribution to the Literature

Since Martin and Smith (1999) attempted to model NHS demand, the NHS has undergone significant structural reorganisation, yet excessive waiting times persist. An extensive search of the literature was performed however, as far as the author is aware, no independent econometric studies concerning the effect of waiting times on demand for NHS care using data generated by the current CCG-Trust structure exist. This dissertation represents an overdue econometric analysis of the demand for NHS services with respect to waiting times.

4 Data and Sample

In this section I provide an overview of the data and their source. An explanation of the variables and their construction is presented. Variables names are in parenthesis. Some descriptive statistics are provided.

4.0.1 Data Collection and Description

The NHS standardises, collects and publishes anonymised data from across the healthcare system in England. Monthly panel data, where many units are observed over time, on Consultant-led Referral to Treatment Waiting Times for elective care (waiting times), Incomplete Treatment Pathways (Demand) and other determinants of demand were obtained for 191 CCG geographical in England the period 2017-2019 from the NHS Statistical Work Areas portal. Each CCG is identified by a unique three digit code. During the study period there was some merger activity in which the newly created resultant CCG was given a new identifier. 12 specialities were pooled to produce aggregate data on elective care as a category. I do not analyse the period impacted by

Table 1: 12 Clinical Specialities

Treatment Function
General Surgery
Urology
Trauma & Orthopaedics
Ear, Nose & Throat (ENT)
Ophthalmology
Oral Surgery
Neurosurgery
Plastic Surgery
Cardiothoracic Surgery
General Medicine
Gastroenterology
Cardiology
Dermatology
Thoracic Medicine
Neurology
Rheumatology
Geriatric Medicine
Gynaecology
Other

Source: NHS, 2020

Coronovirus-19 as discussed.

The data are aggregated at the CCG level. Ideally, this study would use finer granularity 'small' or 'ward' level areas (of which there are approximately 35,000) , or better still, individual patient records to ascertain the effect of waiting times on demand, because information is lost in the aggregation; namely the inter-reigional heterogeneity of waiting lists. Such data, which can be found in the Hospital Episode Statistics database was not available to me in the course of this dissertation due to patient confidentiality reasons, which I am powerless to change. As a result, there may be apparent relationships at the aggregate level which do not exist at the smaller and individual level. I discuss this in detail in the Results section.

The panel is considered 'short and wide' given that whist $T = 36$ is not small, it is smaller than $N = 191$ (Griffiths et al, 2012). On inspection prior to modelling, the data were found to be balanced, characteristically normal and heteroskedastistic. Visual and statistical tests were satisfactory for mild and severe potential outliers and the possibility of processes not present in the main body of the data that might be of concern. Outliers might be generated by hospitals making data inputting errors, or by changes in IT infrastructure. It is important to not be over-zealous with removing data points (Tukey, 1960) as this leads to the data, by definition, failing to be a true sample. This is of critical importance given the fundamental assumption of regression analysis being the data are from a random sample. We should therefore be careful not to lose information by forcing the data, which risks arriving at inferences that are incorrect. In my tests for outliers I see no observations that are nonsensical, or that are likely the results of error. There were no suspiciously low referral rates nor large fluctuation in referral activity in the period.

5 Included Variables

A baseline model of NHS healthcare utilisation analogous to the standard economic model of demand, where price varies inversely with quantity demanded, was considered to include the following explanatory variables, based on the literature explored.

5.1 The Demand For NHS Elective Care

The dependent variable in the model is the rate of demand utilisation `(DEMAND)` for NHS elective care services. These are patients who have been referred by their GP for specialist consideration and are on a Referral-to-Treatment pathway and have not started treatment at the end of the reporting period. They are at various stages of the pathway, they may be waiting for an appointment with a consultant, diagnostic tests, or admission for treatment. For these patients, a clinical decision to admit has been made and the patient is awaiting admission. This volume of Referral-to-Treatment pathways is referred to as the size of the waiting list. These times are aggregated at the CCG level. Previously in the literature they have been aggregated with lower granularity, at the synthetic ward level (Martin and Smith, 1999), where each ward has an average population of approximately 10,000 and at a higher granularity at the regional level (Lindsay and Feigenbaum, 1984) who aggregate data for 14 regional areas.

5.2 Hospital Waiting Times

Similarly, the length of time that an individual has been on a pathway is referred to as the waiting list waiting time (`WAIT_TIME`). Consultant-led Referral to Treatment Waiting Times are measured in weeks and by consultant speciality, which in this dissertation have been aggregated, as such in the model there is one type of treatment for patients. The data includes both patients whose Referral-to-Treatment pathway clock stopped during the reporting period and those for whom the clock was still running at the end of the reporting period. The data is for the average time accrued spent waiting on the list at the end of the reporting period. Median wait times have been calculated from aggregate data and therefore constitute estimates of average waiting times.

According to economic theory, when the cost (dependant on t) of receiving treatment falls, demand should rise in response. However, another possible outcome is that time spent waiting generates additional demand when patients develop co-morbidities. For example, a patent with an untreated endocrine disorder with time might develop renal disease and later bone mineral density might be diminished, all requiring separate referrals where a single patient is counted multiple times because they are waiting in multiple queues. Vy this logic, as the waiting time grows, the incidence of demand for services increases.

Demand and waiting times do not have a straight forward relationship. Johansen (1983) “closed the gap” (Tor Iversen, 1993) between game theoretic and welfare economics approaches, demonstrating queuing as a non-cooperative game. The implication of this development was that queue formation could no longer be argued to be an exogenously determined probability distribution. The decision as to whether or not to join a queue depends on perceptions of waiting time, which is in turn based on other demanders joining the queue. Thus, waiting times are determined endogenously. This issue of endogeneity must be a consideration later when choosing a model, in order to avoid biased and inconsistent estimates (Wooldridge, 2002). Two-stage least squares and generalised method of moments estimation with instrumental variables are common in the literature as a means of addressing endogeneity.

In this dissertation, I employ an instrumental variables approach to account for the presence of this endogeneity (Semykina and Wooldridge, 2010).

5.3 Average Income

In England, the NHS dominates the market for healthcare, however a significant and growing private sector exists (LaingBuisson, 2019). Around 19 percent of the UK population has access to private medical insurance. If patients are faced with excessive wait times, they might seek private sector treatment which invariably has a much shorter waiting list. In recent years physical access to private inpatient bedshas increased. Two large private