Empirical Investigation of Practice Variation of Physicians in Terms of Use of Diagnostic Test Orders under the Effect of Workload and Physician Characteristics

by

Büşra Ergün-Şahin

A Dissertation Submitted to the
Graduate School of Business
in Partial Fulfillment of the Requirements for
the Degree of

Doctor of Philosophy

in

Business Administration

KOÇ
UNIVERSITY

August 5, 2019

Empirical Investigation of Practice Variation of Physicians in Terms of Use of Diagnostic Test Orders under the Effect of Workload and Physician Characteristics

Koç University

Graduate School of Business

This is to certify that I have examined this copy of a doctoral dissertation by

Büşra Ergün-Şahin

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Assoc. Prof. Evrim Didem Güneş (Advisor)

_____

Assoc. Prof. Ayşe Kocabıyıkoğlu [Co-Advisor]

_____

Prof. Zeynep Akşin Karaesmen

_____

Asst. Prof. Mehmet Gçnen

_____

Asst. Prof. Enis Kayış

_____

Assoc. Prof. Raha Akhavan Tabatabaei

Date: _____

I dedicate this thesis to my sons, Ahmet Yiğit and Ali Kerem. Thank you for being with me all the seconds of my work during the last nine months.

ABSTRACT

Empirical Investigation of Practice Variation of Physicians in Terms of Use of Diagnostic Test Orders under the Effect of Workload and Physician Characteristics

Büşra Ergün-Şahin

Doctor of Philosophy in Business Administration

August 5, 2019

In recent years, there has been concerns about the increasing and potentially unnecessary use of diagnostic tests, which not only increases healthcare expenditures, but may also be harmful for the patients. There are numerous factors affecting the physicians' use of diagnostic test orders. In this thesis, we empirically investigate the influence of workload and physician characteristics such as gender and experience on the diagnostic test ordering behavior of physicians. Data from a public hospital is used in this thesis. In the first study, we define workload in two forms: the unfinished workload, that is, the number of patients waiting to be examined, and the finished workload, that is, the number of patients examined.We hypothesize that physicians order more diagnostic tests at higher unfinished workload since physicians use diagnostic tests as a substitute for time with the patient, and physicians order fewer diagnostic tests at higher finished workload due to fatigue and mental depletion. In order to tests these hypotheses, we employed zero inflated negative binomial regression models and we find evidence supporting our hypotheses. In the second study, we analyze how physicians' characteristics affect test ordering behavior of physicians and distribution of daily load, since previous literature suggests that test ordering behavior is adjusted more by physicians' habits and characteristics than by objective evidence and clinical need (Vinker et al. 2007). Employing negative binomial regression model with random effects, we find that experienced physicians and female physicians order more, and distribute their daily load more evenly. Our findings contribute to the body of knowledge in both healthcare operations management and medical literature by showing state dependency of work content and showing that characteristic of workers are also covariates of work content.

ÖZETÇE

Doktorların Teşhis Amaçlı Test Kullanımının İş Yükü ve Doktor Karakteristiğinin Etkisi Altında Ampirik Olarak Araştırılması

Büşra Ergün-Şahin

İşletme Bölümü, Doktora

5 Ağustos 2019

Son yıllarda, teşhis amaçlı testlerin artan ve muhtemelen gereksiz kullanımı tartışma konusu olmuştur, çünkü bu durum sadece sağlık harcamalarını arttırmakla kalmayarak hastalar için zararlı olabilecek sonuçlara da sebep olabilir. Doktorların teşhis amaçlı test isteme davranışlarını etkileyen çok sayıda faktör vardır. Bu tezde, doktorların iş yükünün ve cinsiyet ve deneyim gibi özelliklerinin doktorların test isteme davranışlarına etkisini bir kamu hastanesinden elde edilen verileri kullanarak ampirik olarak inceledik. İlk çalışmada, iş yükü iki şekilde tanımlanır: bekleyen iş yükü, yani muayene edilmeyi bekleyen hasta sayısı ve bitmiş iş yükü, muayene edilen hasta sayısı. İlk hipotezimiz, doktorların bekleyen hasta sayısı arttıkça kısıtlı zamanları olduğundan dolayı hastadan elde edemedikleri bilgiyi testlerden almak için daha fazla test istedikleri yönünde. İkinci hipotezimiz, doktorların toplam muayenesi biten hasta sayısı arttıkça yorgunlukları sebebiyle daha az test istedikleri yönünde. Bu hipotezleri test etmek için sıfır yığılmalı negatif binom regresyon modeli kullandık ve hipotezlerimizi destekleyen kanıtlar bulduk. İkinci çalışmada, doktorların cinsiyetlerinin ve tecrübe yıllarının test isteme davranışlarını ve iş yüklerinin dağılımlarını nasıl etkilediğini analiz ediyoruz, çünkü literatürde test isteme davranışının objektif kanıtlardan ve klinik ihtiyaçlardan çok doktorların alışkanlıklarına ve özelliklerine göre belirlendiğini öne süren çalışmalar bulunuyor (Vinker ve ark. 2007). Rastgele etkiler ile negatif binom regresyon modelini kullanarak, deneyim yılı fazla olan doktorların ve kadın doktorların daha fazla test istediğini ve günlük iş yüklerini daha eşit şekilde dağıttıkları bulduk. Bulgularımız, hem sağlık operasyonları yönetimi hem de tıp literatüründeki bilgi birikimine çalışanların yaptıkları iş içeriğinin duruma (iş yüküne) bağımlı olduğunu göstererek ve cinsiyet ve deneyim gibi çalışanların özelliklerinin de iş içeriği etkilediğini göstererek katkıda bulunur.

ACKNOWLEDGMENTS

I would like to thank my advisors Assoc. Prof. Evrim Didem Güneş and Assoc. Prof. Ayşe Kocabıyıkoğlu for all their support, contribution and valuable comments on this work. I would also like to thank my thesis supervising committee members Prof. Zeynep Akşin Karaesmen and Asst. Prof. Mehmet Gönen for their useful and constructive guidance.

I would also like to extend my thanks to the physicians and the IT team at the hospital for their assistance and valuable comments on this work. I would like to express my special thanks to Dr. Ahmet Keskin for his collaboration and time as well as for his assistance with requesting the data from the hospital.

I wish to acknowledge the support and understanding of my company throughout my study.

Assistance provided by administrative personnel of Graduate School of Business at Koc University was also appreciated.

Finally, I particularly wish to thank my husband, my parents and sisters, all my family and friends for their support and encouragement throughout my study.

This was a long journey starting at Bilkent University and ending at Koç University. I would like to thank Bilkent University as well.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

AIC	Akaike Information Criterion
ED	Emergency Department
GDP	Gross Domestic Product
ICD	International Classification of Diseases
IV	Instrumental Variable
NBREG	Negative Binomial Regression
PR	Patient Record
VIF	Variance Inflation Factor
XTNBREG	Negative Binomial Regression with Random Effects
ZINB	Zero Inflated Negative Binomial

Chapter 1

INTRODUCTION

The increasing cost of healthcare and increasing share of health expenditure in countries' GDP are among the most prevalent issues emerging from debates on the future of healthcare. One of the biggest opportunities for reducing healthcare costs is avoiding the use of medication, tests, and treatments that do not contribute to the patients' health (Gawande, 2009; Hussey et al., 2009; Berwick and Hackbarth, 2012). Although diagnostic tests, such as CT and MRI scans or blood sampling, can be significant sources of information for reaching a diagnosis, and making a treatment plan, overuse of diagnostic tests is also reported (Fryer and Smellie, 2013; Freeman et al., 2014; Khalifa and Khalid, 2014; Shye et al., 1998). In a meta-analysis, Zhi et al. (2013) report an average rate of 43.9% overutilization of tests, while Miyakis et al. (2006) note 67.9% of test orders in the hospital in their study were unnecessary and the Carter Review, a UK Department of Health commissioned review of pathology services in England, reports that 25% of pathology tests were unnecessary (Fryer and Smellie, 2013). Rosenbaum (2017) also reports that up to 30% of health care spendings are wasted. Therefore, waste of healthcare resources is potentially one of the reasons for increasing health expenditures.

Over-testing can be undesirable not only because of costs, but also because of the cost both to the patient and the system such as the anxiety due to false-positive and false-negative results, wrong interventions, the undervaluation of clinical examination, the overcrowding of laboratories, delays in diagnosis in cases where clinical examination would be sufficient (Miyakis et al., 2006). According to Vinker et al. (2007), 30% to 50% of outpatient appointments lead to laboratory test ordering, and Capilheira and Santos (2006) reported diagnostic tests were requested in 55%

of appointments. In our study which uses the data from 2015 and 2016, almost 70% of patient visits to the outpatient units results with laboratory test ordering. While overuse of medical resources is obvious, cutting down the unneeded care without compromising from the quality is the challenging part of this issue. Rosenbaum (2017) mentions the trade-off between overmedicalizing and oversimplifying and tells that "sometimes less is more, sometimes more is more, and often we just don't know", however waste is visible and deserves attention to eliminate waste safely. There are several factors that affect test ordering behavior of physicians. Whiting et al. (2007) provide a comprehensive review of the different factors discussed in the literature, including, diagnostic and treatment related factors (such as ruling in or ruling out a disease, deciding on an appropriate treatment), patient related factors (such as patient demand and demographics), physician related factors (such as clinical experience and cognitive biases), and policy and organization related factors (such as short return time of tests and ease of access to laboratory). Additionally, Fryer and Smellie (2013) mention systemic reasons such as inability to access previous test results, limitations in hospital IT systems and so forth for the increased use of diagnostic tests, and Franks et al. (2000) state that psychological factors, such as risk aversion, anxiety from uncertainty, fear of malpractice, autonomous and controlled motivation for test ordering, and patient centeredness, have been associated with physician behaviors. Shye et al. (1998) suggest time constraint as a possible reason for test ordering.

The above discussion suggests, understanding factors that influence physicians' diagnostic decision making processes can provide important insights for finding appropriate strategies to change test ordering behavior and so to reduce healthcare costs. This thesis empirically investigates how operational variables and physician characteristics affect test ordering behavior of physicians. Chapter 3 focuses on the workload of physicians as a potential factor affecting the number of diagnostic tests ordered per patient, and Chapter 4 studies the effect of physicians' gender and experience on diagnostic test ordering behavior of physicians as well as distribution of daily load.

Studies related to workload in operations serve a great value for operations management literature and implications. Recent literature shows that workload affects service time (Kc and Terwiesch, 2009; Batt and Terwiesch, 2016; Oliva and Sterman, 2001; Berry-Jaeker and Tucker, 2017), and even content of the work (Freeman et al., 2014, 2017; Tan and Netessine, 2014) in different work environments such as banks, restaurants, call centers. As a consequence of these effects, quality, cost and revenue of related services are affected as well. Workload studies in healthcare has also emerged as a promising area of research with a potential for impacting healthcare practice. In this thesis, we focus on diagnostic test orders as the content of work, since number of diagnostic tests ordered per patient constitutes significant part of diagnostic decision making at outpatient units and the work content of physicians which is expected to be affected by workload similar to the studies in the literature.

In Chapter 3, we distinguish between two types of workload when a given patient is being examined: the unfinished workload, defined as the number of patients waiting to be examined, and the finished workload, defined as the number of patients already examined before a given patient. In the literature, workload is generally defined as the number of people waiting to be served, hence, throughout the thesis, we use “workload" to refer to unfinished workload, and “finished load" for the finished workload. We develop two hypotheses in this chapter. First, when workload increases, physicians notice that there is less time available per patient waiting. Even if the actual time available is sufficient, knowing that many patients are waiting may trigger a feeling of rush. In an emergency department setting, Chan (2018) observes that diagnostic tests can be a substitute for careful questioning of patients to gather information more rapidly, although diagnostic tests take time to complete. Hence, we hypothesize that, when the workload is high, physicians would order more diagnostic tests per patient, because they feel that they don’t have enough time to conduct a detailed examination of the patient. Concurrently, they do not want to sacrifice diagnostic accuracy while rushing to finish the care of all waiting patients. Second, medical examination and clinical decision making are knowledge intensive processes and require high cognitive effort (Laker et al., 2017; Chan, 2018). In knowledge

intensive work environments, higher finished load in a work day affects the decision processes, due to fatigue and mental depletion; which may, in healthcare settings, lead to simplified diagnostic decisions, performance deficits and less engagement with patients (Danziger et al., 2011; Hockey and Earle, 2006; Van der Linden et al., 2003). Furthermore, evidence from the literature shows workers cannot sustain the same performance for the whole day (Kc and Terwiesch, 2009; Van der Linden et al., 2003). Then, we hypothesize that, with increasing finished load, physicians would order fewer diagnostic tests per patient.

We test our hypotheses with data from an internal medicine outpatient unit of a public hospital. Our motivation for choosing an internal medicine outpatient unit is the high use of diagnostic tests in decision making, and the high workload. We used zero-inflated negative binomial (ZINB) model to test our hypotheses, since the number of diagnostic test orders per patient as our dependent variable, has an excess number of zeros, and there is over-dispersion, that is, the variance of this variable is much larger than its mean. This model has been used in the literature for similar situations in order to avoid the underestimation of the variability in the data (e.g. Batt and Terwiesch, 2016). We find strong evidence for both of our hypotheses. Physicians order more diagnostic tests per patient in response to increasing workload and order fewer diagnostic tests per patient in response to increasing finished load. In particular, as workload, increases from its 5th to 95th percentile (i.e.,from 0 to 28 patients), the number of tests ordered per patient increases from 10.5 to 12.8, indicating 22% increase. By contrast, as finished load increases from its 5th to 95th percentile (i.e., from 1 to 48 patients), the number of tests ordered per patient decreases from 12.5 to 10.2, indicating 18% decrease. We also provide robustness checks with alternative models, variables and samples that confirm our results.

This study contributes to the growing literature in operations management that suggests workers are responsive to changes in the work environment and demonstrate adaptive behavior in response to the amount of workload (Boudreau et al., 2003; Kc and Terwiesch, 2012; Tan and Netessine, 2014; Freeman et al., 2017; Berry-Jaeker and Tucker, 2017; see Delasay et al., 2016 for a review of related literature), unlike

the existing literature which assume employees work independent of the state of the system including the workload. In particular, there are also studies in the context of diagnostic test ordering behavior of clinicians (Deo and Jain, 2018; Alizamir et al., 2013; see Section 3.2 for a detailed literature review). Our study differs from these studies by suggesting diagnostic test orders as a substitute for time with the patient, and investigating the effect of workload and finished load on the content of work, rather than service time. Furthermore, we use finished load as a proxy for fatigue in this empirical study, while studies of fatigue in prior medical literature rely on self-reported measures (Mazur et al., 2016; Gaba and Howard, 2002). Our results suggest that operational interventions that smooth the workload at polyclinics have the potential to reduce such unnecessary care, which is discussed further in Section 3.8.

In Chapter 4, we empirically study whether there is association between physicians' experience and gender and their test ordering behavior, since previous literature suggests that test ordering behavior is adjusted more by physicians' habits and characteristics than by objective evidence and clinical need (Vinker et al., 2007). Additionally, we study the association between physicians' characteristics and distribution of daily load. We introduce four hypotheses in this chapter. First, physicians who have been practicing for more years is assumed to have improved skills and deeper clinical knowledge which is expected to lead to superior clinical abilities. This learning by doing mechanism has been found in different industries including healthcare (Reagans et al., 2005). However, medical developments occur continually and treatment standards change with new clinical evidence, and the knowledge that physicians accumulate over the years may easily become out of date. Therefore, it is also probable that physicians with more experience may be less likely to provide clinically appropriate care (Choudhry et al., 2005). Furthermore, education methodology and training practices also changes over the years, and these changes create cohort effects (Tsugawa et al., 2017b). Then, we hypothesize that physicians' years of experience in practice cause practice variation in terms of number of diagnostic test orders and distribution of daily load. Second, studies also found that male and

female physicians have differences in their practice patterns such as female physicians may be more likely to follow evidence-based practice and clinical guidelines (Baumhäkel et al., 2009; Berthold et al., 2008; Kim et al., 2005), perform better on standardized examination (Jerant A, 2013; Roter et al., 2002; Ferguson et al., 2002). Also, female physicians could be affected by the test demand of their patients more than their male colleagues or patients might be more demanding when they are examined by a female physician compared to a male physician, since female physicians engage in more communication with their patients (Roter et al., 2002). Furthermore, studies from other industries indicate that women have more cautious approaches to solve complex problems (Powell and Ansic, 1997; Barber and Odean, 2001; Charness and Gneezy, 2012). Therefore, female physicians could order more diagnostic test orders for their patients and distribute their daily load more evenly due to their cautiousness. So, we hypothesize that female physicians will order more diagnostic tests and distribute their daily load more evenly.

In order to test these hypotheses, we used data from 5 outpatient units (internal medicine, gynecology, pediatrics, hematology, endocrinology) of a public hospital (see Section 4.3 for details). Our motivation for choosing these outpatient units is the high use of diagnostic tests in decision making. We used negative binomial regression (NBREG) with random effects in order to capture individual test ordering behavior of physicians and due to over-dispersion of the number of diagnostic test orders per patient. In our study sample, we have 119,254 patient visits (after some exclusions), which are generated by 42 physicians. We developed another model to analyze the relation between physician characteristics and distribution of daily load. Results of our models indicates that physicians have practice variations in association with their experience and gender. As physicians' experience, defined by years after graduation, increases from 0 to 35 years, the mean number of diagnostic test orders increases from 8.5 to 10. We also find that the mean number of diagnostic test orders for male physicians is 8.6 and for female physicians 9.5, while holding all other variables at their means. We also found that female physicians and experienced physicians distribute their daily load more evenly.

We contribute to this literature by using patient level retrospective data instead of survey data or physician level data which are used mostly in the literature; by using operational variables such as finished and unfinished workload of the physician at the time of patient examination instead of using only physician, patient and hospital related control variables.

In both of the studies, introduced above, our aim is not to analyze whether physicians order adequate number of tests or unnecessary tests, nor to evaluate the impact of requested tests on the patient outcomes or quality of care. Because of increasing use of diagnostic test orders as well as increasing cost of healthcare, we would like to investigate whether mechanisms, such as rush and fatigue, emerging due to operational variables, such as workload and unfinished load, as well as physician characteristics, such as experience and gender, create practice variations. This knowledge can provide important insights for healthcare managers in developing new strategies and policies for the use of healthcare resources, as well as design and implementation of interventions to change test ordering behavior of physicians and to reduce healthcare costs. However, healthcare managers, who work on improving test ordering behavior of physicians, should be aware that multiple mechanisms could be active at the same time and contextual variables also affect physicians' behavior (Van der Weijden et al., 2002).

The next chapters of the thesis is structured as follows. Chapter 2 introduces data collection process. Chapter 3 includes details of the first study with sections of the related literature (3.2), empirical setting (3.3), hypotheses development (3.4), data set and variables (3.5), econometric model and results (3.6), robustness checks and alternative explanations (3.7) and finally discussion and conclusion (3.8). And Chapter 4 includes details of the second study with sections related to literature review (4.2), hypothesis development (4.3), empirical setting (4.4), data set, variables and methods (4.5), results of the model (4.6), further analysis (4.7) and discussion and conclusion (4.8). And then, Chapter 5 concludes the thesis and mentions future research opportunities in accordance with the findings in this thesis.

Chapter 2

DATA COLLECTION

After deciding on the research questions for this thesis, we got in contact with physicians from a large scale public hospital in order to discuss the problem with them and rationalize the problem. When the research questions matured, we presented our research proposal to the ethics committee of the hospital in association with a physician from the hospital. After getting approval of the ethics committee, we conducted field observation at the outpatient units and observed whole process from registration to the end of examination. We have been among patients at the waiting area and with physicians during examination. Appendix A includes photos from the field observation including patient registration at registration booth, screens to follow sequence at waiting area and physicians' computer screen showing registered patients and electronic medical records page. After understanding whole process and how the data is collected, we requested retrospective data from IT department of the hospital in accordance with our research question. Details of the patient flow are mentioned in Section 3.3.

The whole data set includes patient records from five different outpatient units (internal medicine, gynecology, pediatrics, hematology, endocrinology) of the hospital and its district polyclinics collected in 2015 and 2016. For each patient record, the data included information on (1) patient gender, (2) patient age, (3) patient id number, (4) the attending physician, (5) gender of the physician, (6) experience of the physician, (7) the ICD-10 code, (8) the registration time, (9) the examination time, (10) the test ordering time, (11) the number of diagnostic tests ordered, (10) the types of tests ordered, (12) the cost of diagnostic tests ordered and (13) patient's total number of visits to the hospital, (14) the type of outpatient unit, (15) location. After examining the data set, we decided to use the data from internal medicine of

the hospital for the first study presented in Chapter 3 and the whole data of 2016 for the second study presented in Chapter 4. Details about each data set are explained in related chapters.

Before starting our analysis, we examined the raw data and made data cleaning to get our study sets. First, we excluded patient records with irregular flow such as examination time is recorded after diagnostic test order time and incomplete patient records from study sample. Second, we excluded patient visits that are recorded out of working hours, i.e., before 8am and before 5pm. Further data cleaning in accordance with the research questions is conducted and mentioned in related chapters. Although we made some data cleaning, we considered excluded patient records while making workload analysis.

Chapter 3

HOW DOES WORKLOAD AFFECT TEST ORDERING BEHAVIOR OF PHYSICIANS? AN EMPIRICAL INVESTIGATION

3.1 Introduction

In this chapter, we focus on the workload of physicians as a potential factor affecting the number of diagnostic tests ordered per patient. We distinguish between two types of workload when a given patient is being examined: the *unfinished workload*, defined as the number of patients waiting to be examined, and the *finished workload*, defined as the number of patients already examined before the patient. In the literature, workload is generally defined as the number of people waiting to be served, hence, throughout the paper, we will use “workload" to refer to unfinished workload, and “finished load" for the finished workload.

When workload increases, physicians become aware that there is less time available per patient waiting. Even if the actual time available is sufficient, knowing that many patients are waiting may also increase their stress level and induce a feeling of rush. In an emergency department setting, Chan (2018) observes that, although formal tests take time to complete, and can thus prolong the length of stay, testing can also be a substitute for careful questioning or serial monitoring to gather information more rapidly. Hence, we hypothesize that, when the workload is high, physicians would order more diagnostic tests per patient, because they feel that they don’t have enough time to conduct a thorough examination of the patient. Concurrently, they would like to get more information about the patient to increase diagnostic accuracy, while rushing to finish the care of all waiting patients.

Medical examination and clinical decision making are knowledge intensive pro-

cesses and require high cognitive effort (Laker et al., 2017; Chan, 2018). In knowledge intensive work environments such as courts (Danziger et al., 2011), higher finished load in a work day affects the decision processes, due to fatigue and mental depletion; which may, in healthcare settings, lead to simplifying diagnostic decisions, performance deficits and less engagement with patients (Danziger et al., 2011; Hockey and Earle, 2006; Van der Linden et al., 2003). Furthermore, evidence from the literature shows workers cannot sustain the same performance for the whole day (Kc and Terwiesch, 2009; Van der Linden et al., 2003). Hence, we also hypothesize that, with increasing finished load, physicians would order fewer diagnostic tests per patient.

We test our hypotheses with data from an internal medicine outpatient unit in a public hospital. Our motivation for choosing an internal medicine outpatient unit is the high use of diagnostic tests in decision making, and the high workload. Our study uses data for two years (2015 and 2016; 11,271 patient records in total after some exclusions; detailed in Section 3.5) to investigate how workload related mechanisms can affect the test ordering behavior of physicians.

We find strong evidence for both of our hypotheses. Physicians order more diagnostic tests per patient in response to increasing workload and order fewer diagnostic tests per patient in response to increasing finished load. In particular, as workload, defined by the number of patients waiting, increases from 5th to 95th percentile (i.e., from 0 to 28 patients), the number of tests ordered per patient increases from 10.5 to 12.8, indicating 22% increase. By contrast, as finished load increases from 5th to 95th percentile (i.e., from 1 to 48 patients), the number of tests ordered per patient decreases from 12.5 to 10.2, indicating 18% decrease. We also provide robustness checks with alternative models, variables and samples that confirm our results.

Our study contributes to the growing literature in operations management that suggests, unlike most analytical studies which assume employees work at the same rate and provide the same work content regardless of load, workers are responsive to changes in the work environment and demonstrate adaptive behavior in response to the amount of workload (Boudreau et al., 2003, Kc and Terwiesch, 2012,Tan and Netessine, 2014, Freeman et al., 2017, Berry-Jaeker and Tucker, 2017; see Delasay

et al., 2016 for a review of related literature). In particular, there are also studies in the context of diagnostic test ordering behavior of clinicians. Our study differs from other studies in the healthcare literature that study test ordering behavior (Deo and Jain, 2018; Alizamir et al., 2013; see Section 3.2 for a detailed literature review) by positing diagnostic test orders as a substitute for time with the patient, and investigating the effect of workload and finished load on the content of work, rather than service time. Furthermore, studies of fatigue in prior medical literature rely on self-reported measures (Mazur et al., 2016; Gaba and Howard, 2002) in experimental settings, while we use finished load as a proxy for fatigue in an empirical study. This study is also relevant to the unnecessary care phenomenon in healthcare, since changes in the work content in response to workload may result in overtreatment or undertreatment of patients. Our results suggest that operational interventions that smooth the workload at polyclinics have the potential to reduce such unnecessary care, which is discussed further in Section 3.8.

In this chapter, we first review the related literature. The details of our empirical setting are given in Section 3.3. Section 3.4 presents our hypotheses. We provide details of our data set, and the variables used in our analysis in Section 3.5. Section 3.6 presents our results. Robustness checks and alternative explanations are discussed in Section 3.7. Section 3.8 concludes the chapter.

3.2 Literature Review

The study presented in this chapter is related to the growing literature that investigates the effects of workload on server behavior in service systems. These studies focus on the impact of workload on two distinct measures: (i) the content of the work, and (ii) the speed of the workers. These two effects are observed due to different mechanisms, as summarized in Table 3.1. For a comprehensive review of the literature on workload effects in operations management, we refer the reader to Delasay et al. (2016). Since our work contributes to the body of literature which empirically examines how human behavior is state dependent, that is, changes with the effect of operational conditions such as workload, in the following, we provide a

review of the work which is directly related to our research questions.

Previous research has shown that, in service settings, such as hospitals, restaurants, call centers, banks etc., a worker can respond to higher workload (measured by the number of people waiting to be served), by either increasing or decreasing the amount of work they perform per customer and/or the average processing time, which may further impact service quality (Oliva and Sterman, 2001; Hopp et al., 2007; Kc and Terwiesch, 2009; Tan and Netessine, 2014). Several studies investigate the relationship between the processing time and workload in healthcare settings (Batt and Terwiesch, 2016; Deo and Jain, 2018, Kc and Terwiesch, 2009; Kc and Terwiesch, 2012), and show that, speed can have an increasing, decreasing, or non-monotonic relationship with workload. In particular, Kc and Terwiesch (2009) find that transporters at a surgery unit work faster with higher levels of workload, but since this is not sustainable for long time, they slow down again; in other words, they respond to workload non-monotonically. On the other hand, Berry-Jaeker and Tucker (2017) find that the length of stay of patients in inpatient units follows an N shaped pattern with respect to occupancy levels; the length of stay first increases, then decreases, and then increases again in response to increasing workload. It should also be noted that, even when the number of tasks per customer is fixed, and the work is standardized, workload still influences processing times, as a worker can complete a specific set of tasks by working faster or slower (Schultz et al., 1998; Schultz et al., 1999).

Our study differs from the work reviewed above, since we focus on the effect of workload on the content of work, rather than service time. In particular, we focus on the diagnostic test ordering behavior of physicians, which is a discretionary aspect of examination. The literature on discretionary tasks in healthcare settings show that both increasing and decreasing effects of workload on diagnostic test orders may exist and the results are driven mainly by the assumptions on the nature of their relationship with service time. Chan (2018) finds that the number of diagnostic test orders may increase or decrease with workload, depending on whether formal diagnostic tests and treatments are net substitutes or complements of clin-

ical observation and reasoning over time. Deo and Jain (2018) hypothesize, in a tertiary care outpatient setting (opthalmology), physicians decrease test ordering due to an increase in anticipated workload in order to speed up, and in the later part of the day; in that setting, the patient returns to the examination room with the test results in the same day before they can finally check-out (after clinical diagnosis, prescription of medication and/or advises for surgical intervention). However, they cannot find supporting evidence for their hypothesis, concluding that increasing workload cannot be attributed to discretionary task reduction. Alizamir et al. (2013) develop an analytical model for the optimal work stopping policies for the diagnosis process, where the diagnostic tests provide more information and improve accuracy of the diagnosis, while increasing service time. Like Deo and Jain (2018), in their setting, the service is completed when the diagnosis is made after receiving the test results. They find that the optimal region of parameters that makes ordering more tests optimal is an interval which shrinks (i.e. the number of tests ordered should decrease with congestion). In both Alizamir et al. (2013)'s and Deo and Jain (2018)'s settings, since the test results are needed to complete the service, increasing workload makes ordering tests less attractive, because it increases service time. In our setting, on the other hand, a patient is considered served when diagnostic tests are ordered; hence, high workload makes ordering diagnostic tests more attractive, since they are used as a substitute for time with the patient.

Similar reduction in discretionary tasks due to workload has been reported in other settings in the literature. Oliva and Sterman (2001) introduce, in the context of the banking industry, the cutting corners approach, to explain a reduction in service value by reduction of discretionary tasks when workload pressure increases. Hopp et al. (2007) argue that optimal work content per customer decreases as the workload increases in operations systems with discretionary task completion. Freeman et al. (2014) find that higher workload results in reduction of discretionary and resource-intensive interventions such as pain relief in a maternity unit. Tan and Netessine (2014) show that waiters spend more effort to increase sales at lower workloads, and more effort to work faster at higher workloads.

Table 3.1: Related literature

Workload affects content of the work			Workload affects speed of the worker / service time			Multiple mechanisms emerges due to workload	
Paper	**Effect**	**Direction**	**Paper**	**Effect**	**Direction**	**Paper**	**Mechanism**
Freeman et al., 2014	maternity, resource intensive interventions	↓	Kc and Terwiesch, 2009	surgery unit, hospital transporters	↑↓	Batt and Terwiesch, 2016	ED, early task initiation
Deo and Jain, 2018	outpatient units, test orders	–	Batt and Terwiesch, 2016	ED, waiting census	↑	Berry-Jaeker and Tucker, 2012	workload smooting
Freeman et al., 2017	maternity, discretionary services	↓	Batt and Terwiesch, 2016	ED, treatment time	↑↓	Tan and Netessine, 2014	rush, engagement
Hopp et al., 2007	manufacturing, discretionary task completion	↓	Deo and Jain, 2018	outpatient units, work speed	↑	Kc and Terwiesch, 2009	fatigue
Oliva and Sterman, 2001	banking, cutting corners, quality of work	↓	Oliva and Sterman, 2001	banking, cutting corners, time per customer	↓	Chan, 2018	slacking off
Tan and Netessine, 2014	restaurant, sales	↑↓	Berry-Jaeker and Tucker, 2017	inpatient units, occupancy level	↑	Danziger et al., 2011	fatigue and mental depletion
Alizamir et al., 2013	diagnostic services, additional tests	↓	Berry-Jaeker and Tucker, 2017	inpatient units, LOS	↑↓↑	Hockey and Earle, 2006	mental fatigue and task performance
Kuntz et al., 2015	hospitals, discretionary task	↓	Kc and Terwiesch, 2012	ICU, occupancy level	↑	Delasay et al., 2016	literature review
Batt and Terwiesch, 2016	ED, triage test orders	↑	Kc and Terwiesch, 2012	ICU, LOS	↓	Berry-Jaeker et al., 2013	quality concern
Berry-Jaeker and Tucker, 2012	ED, test orders	↑	Berry-Jaeker and Tucker, 2012	ED, processing time	↑		
			Schultz et al.,1999	assembly line	↑		

There are also studies that find different mechanisms emerging due to workload, which may affect diagnostic test orders. Batt and Terwiesch (2016) propose "early task initiation", which leads to more diagnostic testing at triage in response to increasing census of waiting room in ED environment, with a motivation to decrease total service time via decreasing test ordering by physicians. Berry-Jaeker and Tucker (2012) suggest another reason for increasing test orders in ED; workers increase their processing time by performing additional tests on the patient because they do not want any additional incoming patients. In our outpatient context, there is no triage process, and physicians must serve all of the waiting patients before the end of the day; therefore, early task initiation or a motivation to avoid additional patients are not likely to have an impact on test ordering behavior. Furthermore, different than the studies showing the effect of workload in the work environments, such as ED, maternity unit, surgery unit etc., where patient care is emergent, we show the effect of workload in outpatient polyclinic, where patient care is not that much emergent.

Our study also investigates the impact of finished load which may create a "fatigue effect" on server behavior. Fatigue has long been recognized as having a central explanatory role in human performance; it emerges due to extended or demanding mental work, causing reduced work effectiveness and tendency to use effort on subsequent tasks (Holding, 1983). In the healthcare setting, Kc and Terwiesch (2009) study the effect of fatigue on service time of patient transporters. Other empirical research has shown that increased workload increases efficiency until workers get overwhelmed in different service environments, such as customer services, help desks and restaurants (Oliva and Sterman, 2001; Tan and Netessine, 2014). In a study of judges' parole decisions, Danziger et al. (2011) conclude that repeated judgements can increase the likelihood of judges' simplifying their decisions and giving an unfavorable decision later in a day or just before the food break, compared to the very beginning of a day, due to mental fatigue. Parallel to this stream of literature, we show that finished load effects the content of the work, where mental fatigue arising after repetitive examinations can lead to simplifying the work and not ordering as

many tests as the times when few patients have been examined yet.

3.3 Empirical Setting

The setting of this study is the internal medicine outpatient unit (polyclinic) of a large public training and research hospital. Due to reasons detailed in Section 3.5, our analysis focuses on the process before the lunch break.

The working hours at the outpatient unit is from 8:30 to 17:30; however, patients start arriving before 8:30, since they need to take a number from the ticket dispensing machine for examination. This point is timestamped as *registration time* in patient records (PRs). Almost all patients are walk-ins in this setting, and they are ordered according to their registration time (i.e., on a first come first served basis), except for the elderly (65+ years old) and disabled[1]. Hence, there are always some patients registered in the system before 8:30 and there is usually a considerable number of patients waiting especially early in the morning (with a mean of 16 and maximum of 51 patients). Patients have the option to choose among the available physicians at the unit, and they can see the queue length for each physician before they make their choice. On a typical day, there are three physicians at the outpatient unit; the head of the department allocates physicians among the outpatient unit and the inpatient wards. The patients can indicate whether they come for examination or to show test results when they take their registration number; these are two separate queues. Furthermore, if the patient has been to the outpatient unit in the preceding 10 days, they are recorded as a control patient (irrespective of whether they have come for an examination or to show their test results); if the patient has not been to the unit in the preceding 10 days, they are recorded as a new patient. In our study

[1]The data does not specify whether a patient has appointment or not. If a patient has appointment, he/she also takes a ticket upon arrival, hence a registration time is generated, and these patients are also counted in the workload. Physician may give priority to patients with appointments, thus having an appointment may decrease patient's waiting time, while we believe the effect of workload on the test ordering behavior does not depend on the ratio of patients with appointments.

sample, 12% of patients were control patients. However, although not specified in the data, we observed in the clinic that percentage of test consultations are very low in the morning session since hospital administration urges patients to prefer the afternoon hours for test result consultation. Physicians also give low priority to these patients in the morning session.

After registering, patients proceed to the waiting area. In the waiting area, the sequence number of the patients currently being examined by each physician is displayed on a screen. The physicians control the numerator at the waiting room from their computers, and can see the list of patients waiting for them on their screen. At any point in time, the number of patients waiting for a given physician ranges from 0 to 52 with a mean of 12.

When a patient's sequence number comes up, they proceed to the examination room, and the physician opens the record of the patient on their computer. If the patient is a control patient, a pop-up warning appears on the screen. When the physician begins recording the patient's history on the computer, a unique protocol number (for that particular visit) is assigned to the patient and this point is timestamped as the *examination time*. At the end of the examination, the physician may (a) refer the patient to another outpatient unit for further examination, (b) give consultation or write a prescription, (c) admit the patient to the inpatient unit, or (d) order additional diagnostic tests from the laboratory or the radiology department; and this completes the service. If the physician orders tests, the time they enter this request to the system is timestamped as the *test order time*. The patients come back for a post-test consultation later (but not necessarily on the same day), in which case the process of registration starts again (i.e., they take a new sequence number etc.). Figure 3.1 gives a visual summary of the flow at the outpatient unit. We note that blood tests, such as cholesterol and triglyceride are the most frequently requested tests, followed by urine tests and chest X-rays (see Table B.1 in Appendix A).

The number of sequence numbers per day for each physician is restricted as 50 by the registration machine. An additional 15 registration numbers are allocated

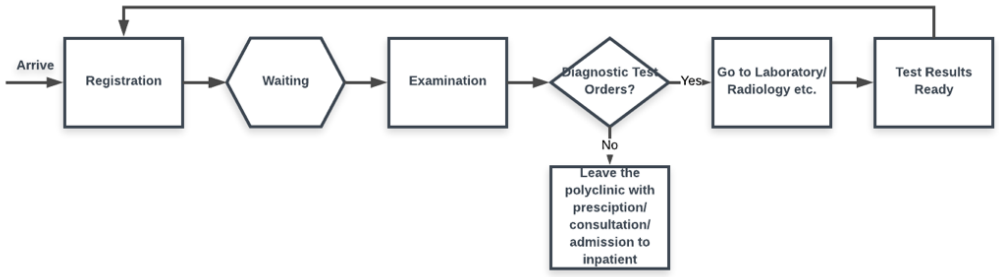


Figure 3.1: Patient flow at the polyclinic

for the elderly and the disabled patients who have examination priority. In special cases, physicians can accept additional patients through the registration desk. The number of patients examined per physician in a given day ranges between 2 and 82 with a mean of 35.4 patients (see Figure 3.2 for a histogram). We note that, only 3% of physician days (39 among 1242) have more than 65 patients. There are also a considerable amount of days, 13.6% (169 among 1242), in which a physician examines fewer than 10 patients. This is observed due to other obligations (such as teaching, inpatient care etc.) of physicians and which prevents them from working for a full day at the outpatient clinic. Note that we restrict our attention to the morning session of four hours, i.e., an average of 35.4 patients, which corresponds to less than 7 minutes per patient. This indicates that most days are highly congested and physicians generally have limited time to examine each patient.

There exists previous literature which suggests physicians' test ordering behavior can be affected by profit concerns (Silverman and Skinner, 2004; Dafny, 2005; Powell et al., 2012). However, the hospital in our study is a public research hospital, and the physicians are salaried civil servants. The total number of patients that they can examine per day is also determined by the hospital management. Physicians are paid based on a fixed salary plus a share from the hospitals' total revenues (in accordance with the average performance of their department and their own performance based on a point system). The number of patients they have seen and the number of diagnostic test orders have a negligible effect on the variance of their total earnings.

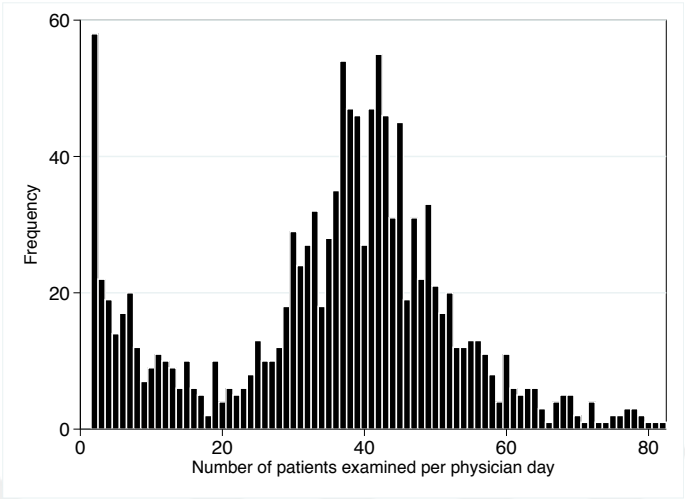


Figure 3.2: Histogram of daily load of physicians

Hence, we can assume financial incentives do not have an effect on the test ordering decisions of physicians in our setting. In addition, even if we assume that some physicians have an incentive to order more tests, that will be a fixed effect and does not interfere with our hypothesis, which will be discussed next.

3.4 Hypotheses Development

In this section, we form the two main hypotheses of the study in this chapter, and relate them to the relevant literature.

3.4.1 Effect of Workload

As mentioned in section 3.3, in our setting, the service episode is completed once (a) the diagnostic tests are ordered, (b) consultancy is given, (c) a prescription is written, or (d) the patient is referred to an inpatient unit or another outpatient unit. The discretionary task, that is, ordering diagnostic tests, helps to increase speed by finalizing the service episode, hence it acts as a substitute for time. This leads us to hypothesize that as workload increases, the number of diagnostic test orders would increase, due to various concerns of the physicians, discussed below.

Physicians, under high workload, may order more tests, to complete the care of

all waiting patients in a limited time; Batt and Terwiesch (2016), Kc and Terwiesch (2009) and Tan and Netessine (2014) find evidence of rushing under increased workload for triage nurses at ED, patient transporters and restaurant waiters, respectively.

While the physicians use their discretion over the process by limiting the time they spend with a patient due to a rush effect to complete the examination of all waiting patients, they would also like to improve diagnostic accuracy under time constraints (Alizamir et al., 2013), and hence, may order more tests to gather more information about the clinical condition of the patient. Workload also triggers a subconscious stress response on workers. Kuntz et al. (2015) show that physicians' tendency to make errors decreases with increasing workload when their workload is low, and increases with increasing workload when their workload is high. In our context, physicians may have concerns about the quality of their clinical decisions when they have time pressure due to workload, and will be more inclined to order tests to increase diagnostic accuracy.

Furthermore, Berry-Jaeker et al. (2013) suggest physicians might use additional tests or treatments to improve their patients' perception of their quality of service. When a service provider (in our case, the physician) performs more tasks for the customer (in our case, the patient), it results in a higher quality experience for the customer (Hopp et al., 2007). Sun et al. (2000) also suggest that ordering more diagnostic tests is a motivation for physicians to increase patient satisfaction. Hence, when physicians spend limited time with their patients due to high workload, they could be inclined to order more diagnostic tests in order to increase patient satisfaction.

Finally, it should also be noted that, since physicians do not have the possibility of denying patients in our context, they may also order tests to create more time with the patient later with more information about patient's health condition, as they will have another encounter for post-test consultation. Of course, this behavior would potentially increase the future workload; however we believe in this context the effect is minimal as post-test consultations constitute a minor part of the workload in the

morning and they are handled mainly in the afternoon session. Reminding that same day post-test consultation patients are not recorded in the afternoon data, 30% of patients are control patients among the patients registered in the afternoon; while 12% of patients are control patients among before noon records as mentioned before (see section 3.5 for further details).

Below, we provide our first hypothesis, which states that, when the number of waiting patients increase (i.e., at high workload levels) physicians will order more diagnostic tests, due to the effects discussed above.

Hypothesis 1: *When the workload (number of patients waiting) at the time of examination of a patient increases, number of tests ordered per patient increases.*

3.4.2 Effect of Finished Load

Our second hypothesis is about the relationship between number of diagnostic test orders, and finished load. Specifically, we hypothesize that the number of diagnostic test orders per patient would decrease with the ordinal position of the patient, or equivalently, the finished load before the examination of the patient. We argue that this is expected due to a fatigue effect.

Previous research suggests, making repeated decisions depletes individuals' mental resources, which can, in turn, influence their succeeding decisions. In an experimental study of office work environment, Hockey and Earle (2006) observe an overall increase in fatigue over two-hour sessions, and find no impairment on primary tasks and decrements for secondary task activities; physicians could consider some of the tests as secondary task for the patient and do not order them due to emerging fatigue. Furthermore, Danziger et al. (2011) show that favorable rulings in judicial decisions decreases with the ordinal position of the cases, because when judges are mentally depleted, they prefer to simplify their decision. Furthermore, service providers can try to minimize the energetical costs of performance by choosing behavioral strategies that require minimal levels of effort (Boksem and Tops, 2008). Medical examination and clinical decision making is an expert service similar to the knowledge intensive environment of judicial decisions, therefore we can expect

fatigue to emerge in this context as well.

To summarize, physicians may not be able to maintain same level of effort throughout the day, and their concerns related to patient satisfaction, malpractice risk and diagnostic accuracy may be diminished by mental depletion and fatigue after examining many patients. This would make them less engaged with the patients and more likely to try to simplify their decision. This may emerge as skipping some of the diagnostic tests that are deemed secondary, or finalizing the examination with the available information. Thus, we state our second hypothesis, below.

Hypothesis 2: *When number of examined patients for a physician during a work day increases, number of tests ordered per patient decreases.*

It should be noted that, fatigue may also emerge as a function of time. It has been shown that level of fatigue can increase after a few hours of task performance (Tanabe and Nishihara, 2004; Hockey and Earle, 2006). However, in our context, time may be associated with other mechanisms that may affect test ordering behavior, and it is difficult to identify the effect of time as a fatigue effect only. Therefore, we chose to include time as a control variable in our models, as opposed to a main effect, since the other mechanisms are beyond the scope of study in this chapter.

Note also that the above hypothesis does not necessarily imply that the quality of the diagnostic decision deteriorates by ordering fewer tests. While that is a possibility, the skipped tests may also be in fact unnecessary. Unfortunately the current data does not allow us to make any conclusions regarding the quality of the decisions, or the necessity of the ordered tests. Hence, in this study we only focus on the number of tests as our dependent variable of interest.

3.5 Data and Variables

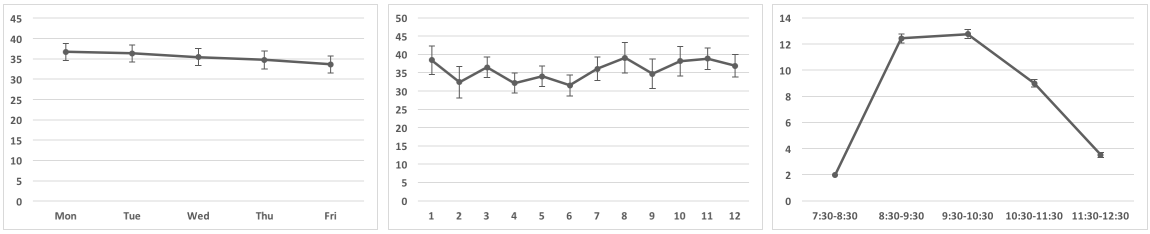
The data for this study is from the outpatient unit described in section 3.3, and encompasses records of 80,367 patient visits for two years, from January 2015 to December 2016. For each patient record (PR), the data included information on (1) patient gender, (2) patient age, (3) patient id number, (4) the attending physician, (5) the ICD-10 code, (6) the registration time, (7) the examination time, (8) the

test ordering time, (9) the number of diagnostic tests ordered, (10) the types of tests ordered, (11) the cost of diagnostic tests ordered and (12) patient's total number of visits to the hospital. Throughout the chapter, we use the individual PRs as the unit of analysis.

From this data, we excluded the PRs for the days where consulting physicians examined patients together with residents for teaching purposes, as then the data does not reflect the actual number of patients seen by the physician. The PRs from physicians who specialize in a subspecialty in internal medicine (such as geriatric, oncology, endocrinology etc.) were also excluded, due to the special needs and treatment plans for these patients. This left us with 6 physicians. To check whether patients with specific complaints have a tendency to choose a particular physician, we conducted chi-square tests for comparing the the proportions of particular ICD codes assigned across physicians. The proportions were not significantly different, leading us to conclude that patients with a specific disease do not have an inclination to choose a particular physician.

We also excluded the PRs after the lunch break since the number of PRs in the afternoon does not reflect the real workload in the afternoon due to practice of using one PR for a patient who visits the clinic more than once in a day (typically patients seen in the morning would come back to show test results in the afternoon). In the remainder of the chapter, we refer to the half day, before noon, as a "day"; hence, "daily" means "before noon". We also excluded the instances where a physician examined only 1 patient a day. After these exclusions, we were left with a sample consisting of 43,966 PRs with 6 physicians and 1,242 physician days; all the graphs and descriptive statistics in the study reflect this data. Figures 3.3a and 3.3b provide a visual summary of the daily average number of patients per day of the week and per month, respectively.

The type and number of tests for different diagnoses can be quite varied; therefore an important control variable is the ICD code, which is expected to be related with the need for testing for a patient. Our data had PRs with 253 different ICD codes. The most frequently used ICD codes are provided in Table B.2 in Appendix A; in



(a) **(b)** **(c)**

*Note: Error bars indicate 95% confidence interval of the mean value.

Figure 3.3: Daily average number of patients examined per physician by (a) day of week, (b) month and (c) time of the day

particular, 74% of the patients were assigned the ICD code Z04.8, which stands for "encounter for examination and observation for other specified reasons". Since this may not reflect the real patient complexity, we excluded the PRs with this ICD code from the study sample (in section 3.7, we provide a robustness check which uses the PRs with the ICD code, Z04.8). In addition, we excluded the ICD codes that were assigned only once during the period represented in the study. This leaves a sample of 11,271 PRs. Note that, all the patients excluded from the analysis sample due to their ICD codes are included in the estimation of the workload measures, since these patients are still examined by the physicians.

We used one hour time intervals in our analysis, since very short intervals increase the amount of unpredictable variability in the variables, and very long time periods limit the number of observations within a day, thus making it difficult to estimate within-day effects (cf. Deo and Jain, 2018). Since we conduct our analysis on the PRs before lunch break, we had five intervals (numbered from 0 to 4) from 7:30 to 12:30. The average number of patients examined per time interval is provided in Figure 3.3c. While the regular working hours start at 8:30, physicians may also start the day early and see a few patients, as reflected in Figure 3.3c.

We also analyzed patient mixes according to assigned ICD codes per time interval in order to check whether a specific type of patient arrives in a specific time period, using chi-square test for proportions. We observed that patient mixes were not

significantly different at different time intervals, except the first one (i.e., between 7:30-8:30), where there are significantly more diabetic patients compared to the other intervals. We confirm that this does not affect our results with several robustness checks in section 3.7, including running the model for each time interval and focusing only on diabetic patients.

Next, we provide an overview of our independent, dependent and control variables.

3.5.1 Independent Variables

The independent variable for our first hypothesis is the physician workload. We measure workload at the server level (cf. Tan and Netessine, 2014) rather than at the organization level (e.g., Kc and Terwiesch, 2009; Kuntz et al., 2015), since test ordering behavior of each individual physician should be affected only by their own workload. We calculate the physicians' workload corresponding to patient i ($waiting_i$) as the number of patients in the waiting room while patient i is being examined by the physician. In the literature, similar measures, such as waiting room census at the ED (Batt and Terwiesch, 2016) or the queue length at assembly work (Schultz et al., 1998) are used to capture workload. In our setting, the majority of patients do not make prior appointments; instead, they register and get their sequence number when they arrive at the hospital. So, the physicians assess their workload from the number of patients waiting. Our observations at the site, and interviews with physicians also confirmed that this is the main load metric that the physicians focus on. Figure 3.4 provides the histogram of the variable $waiting_i$ which varies from 0 to 52 with median value of 11. For our second hypothesis, in order to measure the effect of finished load in a work day on the number of diagnostic tests ordered, we use the number of examined patients before patient i ($finishedload_i$). Figure 3.5 provides the histogram of the variable $finishedload_i$ which varies from 0 to 79 with median value of 19.

We also note the difference between the two independent variables, $waiting_i$ and $finishedload_i$; $waiting_i$ represents the number of patients waiting for the physician

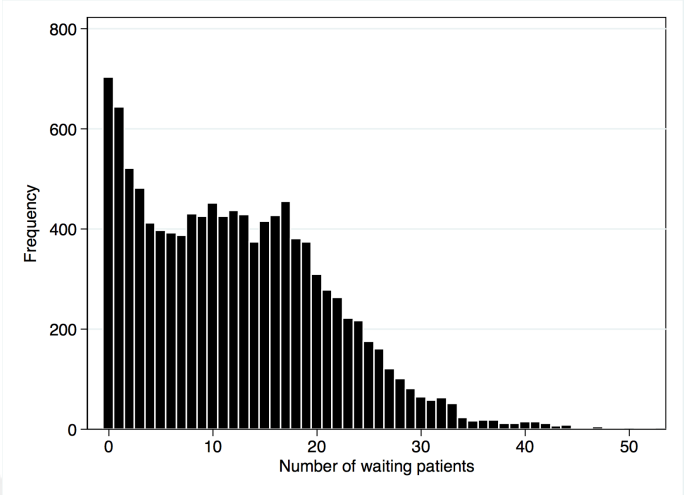


Figure 3.4: Histogram of number of waiting patients

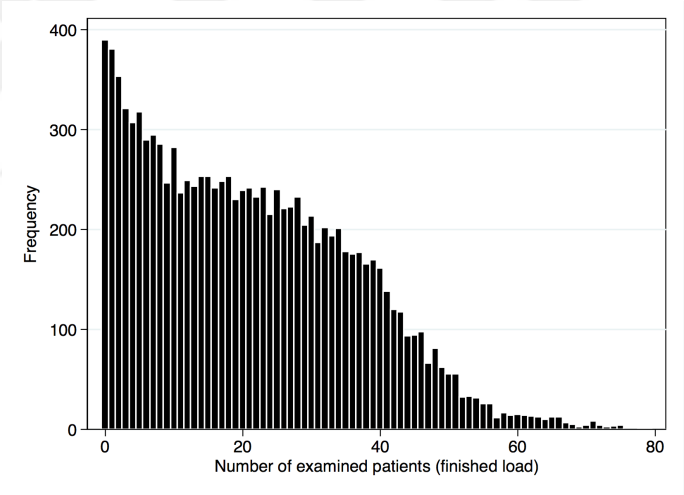


Figure 3.5: Histogram of number of examined patients

when patient i is being examined by the physician, while $finishedload_i$ represents the number of patients examined by the physician up until the examination of patient i (see Figure 3.6).

3.5.2 Dependent Variables

Our dependent variable is the number of diagnostic tests ordered for a patient i by a physician ($ntests_i$). In our study sample, the average number of diagnostic tests

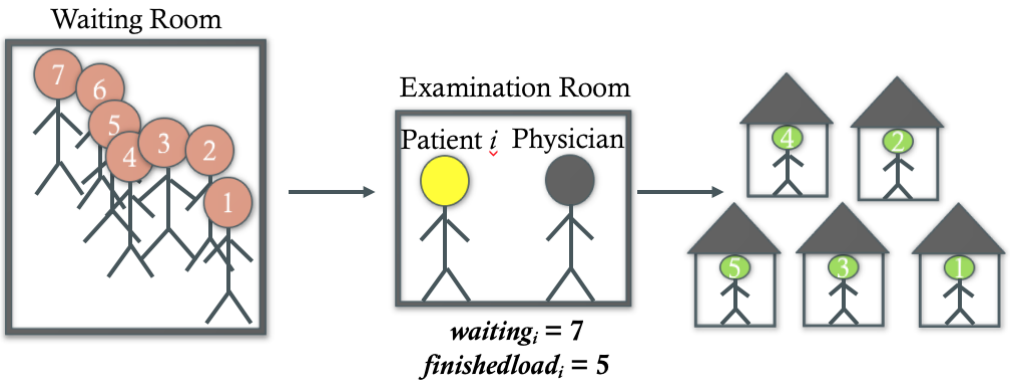


Figure 3.6: Illustration of independent variables waiting and finishedload

ordered per patient is 11.75; Figure 3.7 provides the histogram of the number of test orders. Conditional on at least one test order is given, the average increases to 19.72 tests per patient. Note that each biomarker in a blood test is counted as a new test order, which explains this high figure.

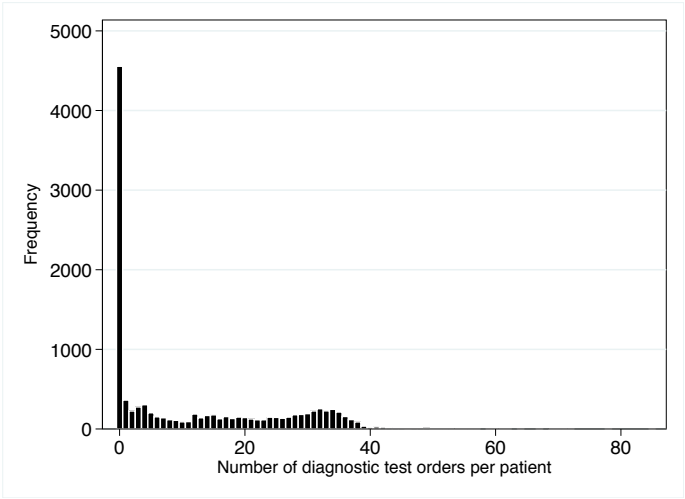


Figure 3.7: Histogram of number of diagnostic test orders per patient

3.5.3 Control Variables

In addition to the variables of interest, we include a number of control variables for each patient i in our study, which can be broadly categorized into: (1) patient related

controls: $status_i$ (control or new patient), ICD_i (since the type and number of test orders for different diagnoses can be quite varied), age_i, $gender_i$, $physician_i$(since test-ordering behavior might differ across physicians), $nvisits_i$ (which indicate the number of patient's previous visits to the hospital, to control for the patient's history at the hospital), and (2) time-related controls: $year_i$, $month_i$,day_i and $time_i$ (time of the day). While $status_i$, ICD_i, $gender_i$, $physician_i$, $year_i$, $month_i$, day_i are categorical variables, the rest are numerical variables.

Among the time related controls we believe the time of day, or the elapsed time since the beginning of the workday, deserves special attention. As discussed above, we used one-hour time intervals to define this variable, taking values from 0 to 4 corresponding to intervals with starting times of 7:30 to 12:30. The elapsed time is expected to be associated with fewer number of diagnostic test orders due to several reasons. Increasing fatigue, as discussed in Hypothesis 2 is one of the possible reasons. Moreover, some patients who know that they will need tests may be arriving earlier in the day. For example diabetic patients need to be fasting during the test and in some polyclinics they are allocated the morning slots. While there is no such allocation in the polyclinic studied, there may be an effect of elapsed time on number of tests due to patients' arriving patterns that may be affected from unobserved clinical needs. Another possible effect is the expected time of test results given the test order time. Tests ordered in the morning will be more likely to be completed within the same day, which may increase the motivation of the physicians to order tests in the morning. Due to all these reasons, we expect a negative effect of elapsed time on the number of tests ordered, and it is important to control for it in testing the hypothesis.

Note that, the variables $finishedload_i$ discussed above, is also a control variable when testing Hypothesis 1, on the relationship between the number of diagnostic tests ordered and workload. Similarly, number of patients in the waiting room while patient i is being examined by the physician ($waiting_i$) is a control variable in the analysis of Hypothesis 2.

Table 3.2 provides descriptive statistics for our dependent and independent vari-

ables for the reduced dataset of 11,271 patient records. These PRs were generated by 9,321 unique patients. 63% of the patients were female and the average age was approximately 48. Note that we drop the subscript from the variables for ease of exposition.

Table 3.2: Summary statistics of dependent and independent variables

Variable	mean(Std.Err.)	min	max	median
waiting	12.21 (0.0832)	0	52	11
finishedload	21.10 (0.1418)	0	79	19
ntests	11.75 (0.1292)	0	85	4

3.6 Econometric Models and Results

The challenge we have in identification of the model is that there is no random assignment of patients to physicians, and patients can actually choose a physician at the registration. This could raise concerns for endogeneity in the workload measures, if there was a reason to believe that test ordering behavior of physicians affect patients' choices. However we find that this is actually not a serious concern, and we will address this issue using an instrumental variables approach later, in section 3.7. In this section the base model and its results are presented first and all the robustness checks are deferred to section 3.7.

3.6.1 Model Formulation

We use a zero-inflated negative binomial (ZINB) model to test our hypotheses, since our dependent variable, the number of diagnostic test orders, has an excess number of zeros (see Figure 3.7), and there is over-dispersion, that is, the variance of this variable (188) is much larger than its mean (11.75). This model has been used in the literature for similar situations in order to avoid the underestimation of the variability in the data (e.g., Batt and Terwiesch, 2016). The ZINB model combines a binary logit process $f_1(.)$, and a negative binomial count process $f_2(.)$, to create

the combined probability density $f(y)$. In this model, "if the binary process takes on a value of 0, with a probability of $f_1(0)$, then $y = 0$; whereas if the binary process takes on a value of 1, with a probability of $f_1(1)$, then y takes on count values 0,1,2,..., from the count density $f_2(.)$" (Cameron and Trivedi, 2009). Then, zero counts are generated in two ways: as the binary process is realized and as the count process is realized, when the binary random variable is 1. It follows that, the ZINB model has a density of

$$f(y|x) = \begin{cases} f_1(1|x_1) + (1 - f_1(1|x_1))f_2(0|x_2) & if y = 0, \\ \{1 - f_1(1|x_1)\}f_2(y|x_2) & if y \geq 1. \end{cases} \tag{3.1}$$

This model has the conditional mean

$$E[y|x] = \frac{1}{1 + exp(x_1\beta_1)} \times exp(x_2\beta_2). \tag{3.2}$$

The vectors x_1 and x_2 in equations (3.1) and (3.2) represent the independent variables to be included in the regression model for the logit, and the negative binomial, parts of the model respectively, and the vectors β_1 and β_2 are the coefficients to be estimated. While the vectors x_1 and x_2 do not need to be the same, we included all the independent and control variables discussed in section 3.5 in both parts of the regression model in our model specification, except for the variable ICD (i.e., the ICD code for the patient); this was due to software limitations. Since our model exceeds the variable limits, we used ICD as a control variable only in the negative binomial part, in order to reduce the complexity of the model. In the logit part of the model, for β_1, positive coefficients indicate an increase in the probability of a zero outcome; hence, a positive coefficient indicates that the dependent variable decreases in the given independent variable. In the negative binomial part, for β_2, positive coefficients indicate an increase in the mean outcome. Because of the two-part and nonlinear nature of the ZINB model, direct interpretation of the coefficients is difficult, therefore, we also report the mean marginal effect of the variables of interest.

The linear predictors $x_1\beta_1$and $x_2\beta_2$ are formulated as follows:

$$x_{i,j}\beta_j = \alpha_j + \beta_{j,1} waiting_i + \beta_{j,2} finishedload_i + \theta_j Y_{i,j} + \gamma_j Z_{i,j} \quad for \quad j = 1, 2 \tag{3.3}$$

where, recall, $waiting_i$ denotes the number of patients in the waiting room while patient i is being examined by the physician, $finishedload_i$ is the number of patients the physician has examined before patient i, $Y_{i,j}$ denotes the vector for patient specific control variables (i.e., $status_i$, ICD_i, age_i, $gender_i$, $nvisits_i$ and $physician_i$) with corresponding coefficient vector θ_j, and $Z_{i.j}$ is a vector of other time related control variables (i.e., $time_i$, $year_i$, $month_i$ and day_i) with corresponding coefficient vector γ_j, and the subscripts 1 and 2 stand for logit and negative binomial parts of the model, respectively. Since the independent variables for Hypothesis 2 are control variables for Hypothesis 1, and vice versa, we use equation (3.3) to test both of our hypotheses.

Table C.1 in Appendix B provides the correlations between the variables included in the model. In order to investigate the existence of multicollinearity among the three variables which were used as both independent and control variables (i.e., $waiting$, $finishedload$, $time$), we calculated the variance inflation factors (VIF) for these variables; a VIF greater than 5 indicates the possibility of multicollinearity (Montgomery et al., 2001). The VIF values for $waiting$, $finishedload$ and $time$ were 1.64, 2.38 and 2.87, respectively, leading us to conclude that multicollinearity is not a concern for our model. Note that in section 3.7, we also present results with alternative models including simpler models such as simple linear regression, which are consistent with results for the ZINB model.

3.6.2 Results

Table 3.3 reports the estimation results of the ZINB model for our two hypotheses. The coefficients for the other control variables are not reported in Table 3.3 for the sake of simplicity, and are provided in Table C.2 in Appendix B. All of the estimations are done using STATA 14.1.

Recall that, in the ZINB model, a positive coefficient in the inflate (logit) part indicates that the dependent variable decreases in the given independent variable, while a positive coefficient in the negative binomial part indicates an increase in the mean outcome. The mean marginal effects are provided in the last column of Table 3.3, and give direct interpretation of the estimated coefficients.

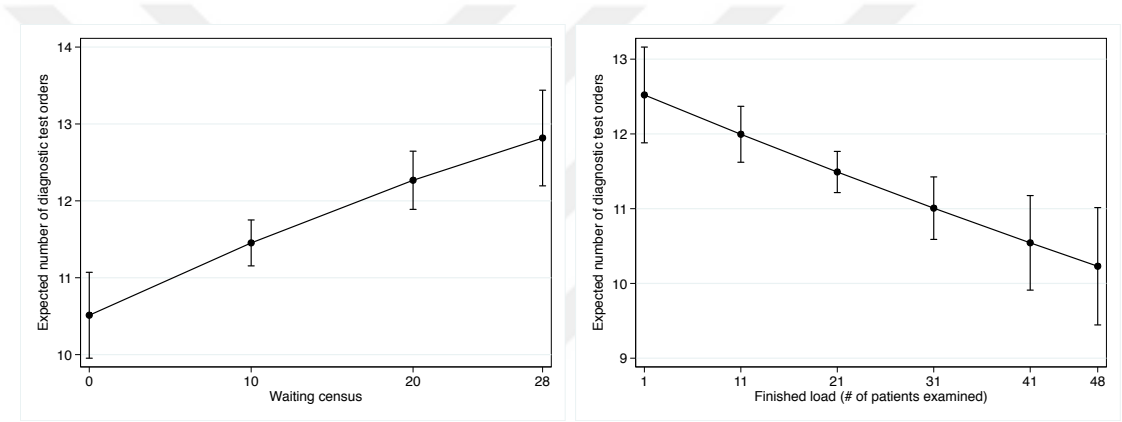
The significant and negative coefficient of *waiting* in the inflation part of the model implies that higher workload, measured by the number of patients waiting to be examined, leads to a decrease in the probability of a patient having a test order of zero. In the negative binomial part, the coefficient of *waiting* is again negative; this implies that the number of tests ordered per patient decreases with increasing workload. In other words, physicians are more inclined to order diagnostic tests for a patient as workload increases, but they order fewer tests per patient. The mean marginal effect of *waiting* is positive and significant, implying, physicians increase test orders by 0.082 tests per patient on average for each additional patient waiting to be examined. Hence, we conclude that, Hypothesis 1 is supported.

Table 3.3: Results from the ZINB model

Variables	Negative Binomial	Inflate (Logit)	Marginal Effects
waiting	-0.002^{*}	-0.029^{***}	0.082^{***}
	(0.001)	(0.003)	
finishedload	-0.004^{***}	0.002	-0.050^{***}
	(0.001)	(0.002)	
Observations		11,271	
	Standard errors in parentheses		
	*** p<0.01, ** p<0.05, * p<0.1		

In order to test Hypothesis 2, we consider the coefficients of the *finishedload*. The positive *finishedload* coefficient in the inflation part of the model suggests, the probability of a patient having zero test orders increases with the cumulative number of examined patients, but is not statistically significant. The negative and significant coefficient for this variable in the negative binomial part of the model implies a

decrease in the number of test orders per patient as the finished load increases. These indicate, although the probability of ordering diagnostic tests is not significantly affected by finished load, number of diagnostic test orders per patient decreases as finished load increases. Consequently, the marginal effect of $finishedload$ is negative and significant, and implies that physicians decrease test orders by 0.050 tests per patient on average for each additional already examined patient, leading us to conclude Hypothesis 2 is supported.



(a) **(b)**

*Note: Error bars indicate 95% confidence interval of the mean value.

Figure 3.8: Expected number of diagnostic test orders according to the model: (a) diagnostic tests vs. waiting census, (b) diagnostic tests vs. finished load

Figure 3.8 presents visual displays of the results discussed in this section, illustrating the magnitude of the effects of the independent variables, from 5th to 95th percentile values with error bars indicating 95% confidence intervals of the mean value. Figure 3.8a illustrates the relationship between the number of diagnostic test orders and workload. The mean number of diagnostic tests per patient increases from 10.5 to 12.8, indicating 22% increase, as the number of patients waiting to be examined varies from 0 to 28. Figure 3.8b indicates that the mean number of diagnostic tests per patient decreases from 12.5 to 10.2, indicating 18% decrease, as the finished load, measured by the cumulative number of patients examined, varies from 1 to 48.

Recall that, in addition to the variables $finishedload$ for Hypothesis 2, and the variable *waiting* for Hypothesis 1, we had patient and time related control variables. We conclude this section with a brief discussion of their coefficients, provided in Table C.2 in Appendix B.

As expected, the marginal effect of the control variable *time* is negative and significant at $\alpha = 0.01$. In fact, we find that time has a strong effect on the number of diagnostic test orders; physicians decrease test orders by -1.512 tests per patient on average for every one unit (i.e., one hour) increase in the time interval. This effect could be investigated further in future research.

The positive and significant marginal effect coefficients of the variables *age* and *gender* suggest physicians order more tests for older patients and female patients. The marginal coefficient of the variable *status* is also positive and significant, indicating, as expected physicians order more diagnostic tests for new patients compared to control patients. Furthermore, the number of diagnostic test orders for a patient decreases with the number of previous visits to the hospital, as indicated by the negative and significant coefficient of the variable *nvisits*. This suggests that previous records of the patient at the hospital is significant for the diagnostic decision process of physicians.

In summary, the results presented in this section support our hypotheses; the content of the physicians' work, measured by the number of diagnostic tests ordered, changes in response to workload, and the finished workload.

3.7 Robustness Checks and Alternative Explanations

In this section, we provide several robustness checks for the results obtained in the previous section, using alternative models, variables and samples, as well as addressing the possibility of endogeneity in waiting.

3.7.1 Robustness Check with Alternative Models

The results presented in section 3.6 were obtained using a ZINB model. We also checked the robustness of our results with different model specifications. To this

end, we replaced our dependent variable $ntests_i$ (which, recall, is defined as the number of diagnostic tests ordered), with a new variable $test_i$, which is a binary variable indicating whether any diagnostic tests are ordered for a patient i, for logistic regression and probit models. The results from three alternative model specifications, specifically, logistic regression, probit and simple regression models are presented in Table 3.4, and confirm those obtained from the ZINB model reported in Table 3.3. In particular, the logistic regression model implies, for one unit increase in the workload (i.e., the number of waiting patients), we expect to see approximately 2% increase in the odds of the physician ordering tests for a patient (see model 2 in Table 3.4). On the other hand, for one unit increase in finished load, we expect to see less than 0.1% decrease in the odds of ordering tests, but this effect is not significant. The results obtained from the probit model and simple linear regression are similar.

We also employed a negative binomial model, which is also used to handle over-dispersed data, and obtained similar results. The AIC (Akaike information criterion; Akaike 1981) of the negative binomial model (69,576.12) is higher than ZINB model (65,271.59), that is, the ZINB model is preferred over negative binomial model; hence, we do not report results of negative binomial model.

Table 3.4: Results from the alternative models

	Logistic(Coeff.)	Logistic(Odds)	Probit	Linear Regression
Variables	(1)	(2)	(3)	(4)
waiting	0.023***	1.023***	0.013***	0.034**
	(0.003)	(0.003)	(0.002)	(0.017)
finishedload	−0.004	0.996	−0.003*	−0.044***
	(0.002)	(0.002)	(0.001)	(0.012)
Observations	11,210	11,210	11,210	11,271
Standard errors in parentheses				
*** p<0.01, ** p<0.05, * p<0.1				

3.7.2 Robustness Check with Random Effects Model

In our study sample, we have 11,271 patient visits over the period of study, which are generated by 9,321 patients; that is, some patients have multiple records/visits (the average number of visits is 1.2 per patient and the maximum number of visits by a patient is 13). In other words, we have repeated measurements from the same patient, which might possibly be correlated. To check whether this influenced our results, we defined an alternative model specification, in particular, a logit model with random effects, to capture the individual level effects. For this model, the intraclass correlation, denoted by ρ, that is, the proportion of the total variance contributed by the panel-variance component, is equal to 0.1607 (see Table D.1 in Appendix C); in other words, approximately 16% of the variance in the propensity to have a diagnostic test order can be attributed to individuals. The likelihood test for parameter ρ provided in Table C.1 shows that, it is significant and the random effects model is appropriate in this case. However, when we compare the coefficient estimates of the logit model discussed in section 3.7.1 and the random effects logit model (see Table D.2 in Appendix C), we observe that, neither the significance nor the direction of the effects change; that is, the presence of random effects do not change the insights from our model.

3.7.3 Robustness Check for Clustered Robust Standard Errors

Getting accurate model estimates is important for accurate statistical inference, a fundamental component of which is obtaining accurate standard errors (Cameron and Miller, 2015). As it is mentioned in section 3.7.2, patient records in our data set can be clustered by each patient, since some patients have multiple records. Therefore, patient records can be grouped into clusters, where standard errors are uncorrelated across clusters but correlated within clusters. One way to control for clustered errors is using random effects estimators as in section 3.7.2. Although, our initial results do not change in random effects model, we also conduct a robustness check to observe whether possible clustering affect standard errors as well as model estimators. By using vce(robust) and vce(cluster) options for ZINB model

in STATA, we obtained model estimates with heteroskedastic robust standard errors and clustered robust standard errors respectively. Table D.3 in Appendix C show that our model results in section 3.6.2 are not biased due to possible clustered standard errors.

3.7.4 Robustness Check for Possible Endogeneity Bias for Waiting

With our first hypothesis, we test whether the number of diagnostic test orders per patient increases in workload, measured by the number of patients waiting to be examined. However, the number of patients waiting could also increase with the number of diagnostic test orders per patient; a physician ordering more tests could be preferred by more patients. In other words, there may be a potential endogeneity bias in our study, because of the unobserved simultaneity between the independent variable *ntests* and the dependent variable *waiting*. To overcome this potential endogeneity bias, we utilize the Instrumental Variables (IVs) approach. In particular, we introduce two IVs, which change with changes in the potential endogenous regressor (*waiting*), but do not lead to changes in the outcome (*ntests*) (Cameron and Trivedi, 2009): (i) the registration time interval of the patient (*reg_interval*), and (ii) the average number of waiting patients for the same physician, for the same time interval of the same day of week (*avwaiting*). Our first IV, the registration time interval of the patients is negatively correlated with *waiting* (-0.57); as expected, there are fewer patients waiting when a patient who has registered earlier is being examined. The correlation between *waiting* and our second IV, *avwaiting* is 0.67. The correlation between the two IVs and the other variables in the study is provided in Table D.4 in Appendix C.

IVs are variables that have an impact, both theoretically and conceptually, on the suspected endogenous variable, that is, the number of waiting patients; and therefore, appear in equation below (i.e., they are relevant), but do not affect the test ordering behavior, and therefore do not appear in the model equation 3.4 (i.e., they are valid) (Wooldridge, 2009,pp.508-520):

$$ntests_i = \delta_0 + \delta_1 waiting_i + \delta_2 finishedload_i + \psi Y_i + \omega Z_i + \alpha, \tag{3.4}$$

$$waiting_i = v_i + \eta_1 avwaiting_i + \eta_2 reg\_interval_i + \eta_3 finishedload_i + \lambda_1 Y_i + \lambda_2 Z_i. \tag{3.5}$$

where recall, Y_i is the vector for patient specific control variables, and Z_i, for time related control variables.

The consistency of the endogeneity test, as well as the coefficient estimates of the IV approach depend on the relevancy and validity of the instruments; hence, we conducted formal tests for relevancy and validity. Although we used the ZINB model in section 3.6, we employed the linear regression model from section 3.7.1 (see Table 3.4) for the tests for validity and relevancy, since these tests are readily available for linear regression models, ie. ivreg2 in Stata 14.1. Hence, the true critical values of the tests and significance levels of the hypothesis tests for relevancy and validity of IVs may differ from those that are reported here, but are expected to be consistent (cf.Freeman et al., 2017). Table D.5 in the Appendix provides the results. To check for relevancy, we conducted under-identification and weak identification tests. Sanderson-Windmeijer test indicated the null hypothesis of under-identification is rejected ($p = 0$). For weak identification, the Cragg–Donald Wald F statistic is equal to 1002.13, with critical F values 19.93 and 11.59, for maximum biases of 10% and 15%, respectively (see Table D.5). Since the estimated F statistic exceeds the critical F values, we can reject the null hypothesis that the instruments are weak for that level of bias; indicating that there is no evidence of our models being affected by the problem of weak instruments.

After confirming the relevancy of the instruments, we also check if they are valid, that is, if they are (i) uncorrelated with the error term and (ii) correctly excluded from the output equation (i.e., they only indirectly influence dependent variable) (Wooldridge, 2009, pp.508-520). The Sargan-Hansen (SH) test which jointly tests (i) and (ii) is not significant (see Table D.5), indicating the instruments are valid.

These results together indicate there is no evidence to reject that our instrumental variables are relevant and valid.

After confirming our IVs are relevant and valid, we test for the endogeneity of waiting (Wooldridge, 2009, pp.527-528). Hausman specification test does not reject the null hypothesis that the variable is exogenous ($p = 0.5184$), so we conclude that the threat of endogeneity bias due to simultaneity between *ntests* and *waiting* is not a serious concern.

Additionally, we could control for endogeneity by using the structural model approach with two step estimation (Cameron and Trivedi, 2009). Structural equation for count number of diagnostic tests is formulated linearly as in equation 3.4, where we suspect that *waiting* is endogenous variable. Namely, we first estimate Equation 3.5 by OLS, and then generate residuals $\hat{v}_i$. Then, due to the complex nature of the ZINB model, we use and estimate the parameters of the NB model given in equation 3.6 below; note that, this is the same as equation 3.4, except for the additional variable, $\hat{v}_i$.

$$ntests_i = \delta_0 + \delta_1 waiting_i + \delta_2 finishedload_i + \psi Y_i + \omega Z_i + \hat{v}_i + \epsilon. \tag{3.6}$$

If the estimated coefficient of $\hat{v}_i$ is not significantly different than 0, then *waiting* can be treated as exogenous, whereas the estimated coefficient of $\hat{v}_i$ is significantly different than 0, the error term α_i and *waiting* are correlated due to endogeneity (Cameron and Trivedi, 2009). For our negative binomial model, we are not able to reject the null hypothesis that the estimated coefficient of $\hat{v}_i$ is 0 ($p = 0.368$, see Table D.6); hence, there is no evidence for endogeneity bias.

3.7.5 Robustness Check with Alternative Variables

We also checked our results with alternative definitions of our two independent variables.

First, we defined an alternative variable for finished load. Recall, $finishedload_i$ reflects the total number of examined patients before patient i since the beginning of

the day. We replaced this variable with $finload1hour_i$, which is defined as the total number of patients examined in the one-hour interval preceding the examination of patient i. The results are presented in Table D.8 in the Appendix C, and confirm our results. We also note that correlation between $finload1hour$ and $waiting$ (-0.1167) is lower compared to the correlation between $finishedload$ and $waiting$ (-0.4893). Hence, this provides another check to ensure that the results are not affected by this correlation.

We also ran our model with a different workload variable. In particular, when we analyzed the effect of workload ($waiting$) at each one-hour time interval by running separate logit models, we observed that, the effect of workload increases with time (see Table D.9 in Appendix C). This is intuitive; for example, the effect of thirty patients waiting for the physician at the beginning of the day versus at the end of the day could be different, the latter creating more rush feeling. Hence, we defined a new workload variable, $remainingload_i$, which is obtained by dividing the number of waiting patients when patient i is being examined to the remaining time (in minutes) till the lunch break. When we run the ZINB model by using this new load variable, the insights obtained in section 3.6 are preserved (see Table D.10).

We also ran our main ZINB model with an interaction term for $waiting$ and $time$; the results obtained are provided in Table C.10 in the Appendix, and are consistent with the marginal effects of the variables from section 3.6.

Finally, we normalized the variable $waiting_i$, by using the mean and standard deviation of the number of patients waiting for the physician examining patient i. Note that, we have a control variable, $physician_i$; however, each physician examining could be used to a different workload level. When we run the ZINB model by using the normalized variable ($normwaiting$), the results obtained in section 3.6 are preserved (see Table D.12 in Appendix C).

3.7.6 Robustness Check with Alternative Samples

The analysis presented in section 3.6 was done on PRs with ICD codes excluding Z04.8, which, recall, stands for "encounter for examination and observation for other

specified reasons". In this section, we provide estimations of a number of alternative model specifications by using different samples to check the robustness of our results.

We first estimated the ZINB model specified in section 3.6 using three different samples, (1) diabetes patients (identified by ICD code E13.8), (2) hypertension patients (identified by ICD code I10), and (3) patients with ICD code of Z04.8. Note that, diabetes and hypertension patients were included in the sample of the analysis presented in section 3.6, while, patients with ICD codes of Z04.8 were excluded, as stated above. The estimated models are provided in Table D.13 in Appendix C, and confirm the results from Table 3.3.

We also estimated our ZINB model without excluding any ICD codes. However, this model did not converge; hence, we limited the iterations to 200, 300, 500 and 1000. The results from these runs are provided in Table D.14 in Appendix C, and confirm our findings from Table 3.3. Also, the results of runs with 500 and 1000 iterations are very close.

Finally, we conducted our analysis with data from another clinic, specifically, the gynecology polyclinic of the same hospital. This sample covers the same time interval as our original data (i.e., from January 2015 to December 2016), and includes 14 physicians and 55,424 patient visits, with a mean of 3.65 diagnostic test orders per patient. Because of the over-dispersion in data and the excess number of visits with 0 test orders (see Figure D.1 in Appendix.C), the ZINB model is appropriate. The results from this model are also in line with our original results (see Table D.15 in Appendix C).

3.7.7 Further Analysis for the Effect of Time

As it is mentioned previously in section 3.5., fatigue may also emerge as a function of elapsed time besides the finished workload, that is analyzed in this study in detail. We would like to discuss the results of our main model related to the effect of time which requires further attention. In the literature, it has been shown that level of fatigue can increase after a few hours of task performance (Tanabe and Nishihara, 2004; Hockey and Earle, 2006; Boksem and Tops, 2008). Since patient

examination and clinical decision making is a cognitively demanding task(Laker et al., 2017; Chan, 2018), after a few hours of examining patients, physicians may exhibit symptoms of tiredness and fatigue. Furthermore, in a study of judges' parole decisions, Danziger et al. (2011) conclude that repeated judgements can increase the likelihood of judges' simplifying their decisions and giving an unfavorable decision later in a day or just before the food break, compared to the very beginning of a day, due to mental fatigue. To summarize, physicians may not be able to maintain same level of effort throughout the day, and their concerns related to patient satisfaction, malpractice risk and diagnostic accuracy may be diminished by mental depletion and fatigue. Hence, we could also hypothesize that as the elapsed time during the work day of a physician increases, number of tests ordered per patient decreases. However, in our context, time may be associated with other mechanisms that may affect test ordering behavior as it is explained in section 3.5, and it is difficult to identify the effect of time as a fatigue effect only. Therefore, while interpreting the results of our main model (ZINB model in section 3.6), we should consider that in addition to fatigue effect, *time* variable could also include other effects that strengthens the effect of time. In other words, the effect of fatigue due elapsed time may not be as much as we observed in the results of *time* variable estimated by our main model.

In particular, the coefficient of this variable is significant and positive in the inflation part of the model, and significant and negative in the negative binomial part of the model, suggesting an increase in the probability of a patient having zero test orders and a decrease in the number of tests ordered per patient, respectively, as the elapsed time increases. In other words, as the time passes from early morning to noon, both the proportion of patients receiving test orders and the number of tests per patient decreases. Consequently, we observe a negative mean marginal effect for the variable *time*; physicians decrease test orders by -1.512 tests per patient on average for every one unit (i.e., one hour) increase in the time interval (Table C.2). According to predicted results of the model, the mean number of diagnostic tests per patient decreases from 13 to 8.6 while time passes from 8: 30 (interval 1) to 12:30 (interval 4). Additionally, we confirmed the results in Table C.2 by replacing

the elapsed time variable, *time*, which is defined on one-hour time intervals, with a continuous variable, *timecontinuous*, which is defined as the number of minutes that have passed till the examination time of the patient. The corresponding results are presented in Table D.7 in Appendix C, and confirm our results.

Although we are not able to separate the fatigue effect of time from other effects in this study, it deserves further attention due to its effect size. The exact mechanism for the effect of time on number of tests ordered, and the effect of time on the quality of the test ordering decision (whether it is associated with overtesting or undertesting) can be investigated further as a future research.

3.8 Discussion and Conclusion

The study presented in this chapter contributes to the growing body of literature that suggests workers are not state independent and show adaptive behavior in response to changes in the system, such as the amount of workload (Schultz et al., 1998; Schultz et al., 1999; Freeman et al., 2017; Batt and Terwiesch, 2016; Berry-Jaeker and Tucker, 2017). Using data from the internal medicine unit of a public research and training hospital, we find that, the content of the work, that is, the number of diagnostic tests ordered per patient, changes in response to workload. In particular, we find that the number of diagnostic test orders increases with workload, measured by the number of patients waiting to be examined. This might be due to various factors, such as the physicians' need to serve all patients in a limited time, improve diagnostic accuracy or increase patient satisfaction. We also find that, physicians order fewer tests per patient as the finished load, measured by the number of patients already examined increases. Fatigue and mental depletion may dominate concerns like malpractice, diagnostic accuracy, and patient satisfaction, leading physicians to be less engaged with patients, and start ordering fewer tests.

Our results imply that work content is affected by changes in the system in healthcare service environments, which may have operational and quality implications. Test related costs currently constitute around 3-5% of the total healthcare costs (Zhi et al., 2013). Our findings suggest that in addition to the remedies sug-

gested in the literature, such as computerized clinical decision support tools that use evidence based test ordering algorithms, or allowing easier access to previous test results if available (Khalifa and Khalid, 2014), reducing the workload and time constraint of the physician may also be effective in reducing costs. Workload effect may be reduced by interventions to increase available physician time per patient, such as reducing the number of patients allocated to a given physician, or using appointment systems, which would decrease the number waiting in the clinic, thus alleviating the rush effect. In order to represent the appointment system, we calculated the predicted number of diagnostic tests from our base model, i.e., when there is 2 waiting patients, who have appointment in the beginning of the day, instead of 16 registered patients waiting before the physician starts working, as it is mean number of waiting patients currently. We observe that predicted mean number of diagnostic test orders decreases from 14.5 to 13.2, indicating 9% decrease, while number of waiting patient in the beginning of the day (when finished load is 0 as well) decrease from 16 to 2. This case shows that using appointment system have the potential to decrease number of diagnostic tests by 1.3 per patient. Considering total number of patients cared in the polyclinic for whole day (around 100 patients by 3 physicians), the hospital has the opportunity to save the cost of around 130 diagnostic tests per day if they switch to appointment system fully. Global effect of this change on the cost of care seems to be non-negligible, so it deserves attention for further research. Finally, training medical students in establishing and maintaining the patient-physician relationship in the face of time pressure (Dugdale et al., 1999) could be also useful.

In our study, to measure examination duration, we use the difference between the start of a given patient's and the subsequent patient's examination as a proxy. This does not give an exact measure, since we cannot observe what physicians do between patients, and a long time interval between two patients may indicate that the physician was idle, or busy doing other activities during this period. However, this would be a good proxy in busy settings, which is a valid observation for the internal medicine unit in our study.

Our data structure also does not allow us to detect the comorbidities of patients, which would increase the number of diagnostic tests ordered for the patient. However, comorbidities are not expected to be correlated with our independent variables ($waiting$, $finishedload$); hence, the coefficients of our covariates would not be effected by this omission. Furthermore, comorbidities would potentially increase with age, and be more prevalent with particular diagnoses such as hypertension (Liu et al., 2016; Piccirillo et al., 2008), and with patients who visit the hospital more frequently. Since we control for age, ICD code and the number of visits to the hospital, we partially control for comorbidities through these variables. However, we note that, the results related to these variables should be interpreted in the light of the possible bias, since their effects could be upward or downward biased due to effect of comorbidity. Another limitation of our study is the number of physicians that generates the data. After some exclusions, we had 11,271 PRs of 6 physicians, and confirmed our findings using 55,424 PRs of 14 physicians from gynecology polyclinic. These findings should be confirmed in other settings with larger data sets.

Possible extensions of our work might be studying the impact of other factors, such as the cost or return time of tests on physicians' behavior. Furthermore, our data does not allow us to observe whether a test is necessary or not; future work could investigate the issue of necessity of tests and the accuracy of diagnostic decisions in the face of increasing workload to generate further insight about physicians' decision making processes. Recall that the exact mechanism for the effect of time on number of tests ordered, and the effect of time on the quality of the test ordering decision (whether it is associated with overtesting or undertesting) can be investigated further as a future research as well.

Chapter 4

PRACTICE VARIATIONS OF PHYSICIANS: EXPERIENCE AND GENDER EFFECTS

4.1 Introduction

In this chapter, we empirically investigate the effect of physicians gender and experience on practice variation of physicians. First, we study the association between diagnostic test ordering behavior of physicians and physicians' experience and gender, since there exists previous literature which suggests that test ordering behavior is adjusted more by physicians' habits and characteristics than by objective evidence and clinical need (Vinker et al., 2007). In addition to practice variations in terms of number of diagnostic test orders, we also analyzed how physicians distribute their daily load and whether practice variation exists in response to physicians' gender and experience. In Chapter 3, we find that number of diagnostic test orders per patient increases with increasing load of physicians due to rushing, since physicians use diagnostic tests as a substitute for time with the patient. In this chapter, we investigate the same effect in addition to the effect of physicians' gender and experience on diagnostic test ordering behavior. In relation to these, we analyze how physicians distribute their daily load in order to observe the rush effect, and then we analyze whether physicians' gender and experience are associated with this distribution.

First, physicians who have been practicing for more years is assumed to accumulate tacit knowledge and improved skills which is expected to lead to superior clinical abilities; this learning by doing mechanism has been found in different industries including manufacturing, service settings and healthcare (Reagans et al., 2005). However, medical advances occur frequently and treatment standards change with new clinical evidence, and the explicit knowledge that physicians possess over the

years may easily become out of date. Then, we hypothesize that physicians' years of experience in practice cause practice variation in terms of number of diagnostic test orders and distribution of daily load. However, the effect of physicians' experience on the number of diagnostic test orders and distribution of daily load could be in two ways due to the factors discussed above.

Second, studies also found that male and female physicians have differences in their practice patterns such as female physicians may be more likely to follow evidence-based practice and clinical guidelines (Baumhäkel et al., 2009; Berthold et al., 2008; Kim et al., 2005), perform better on standardized examination and have better communication and counseling skills (Jerant A, 2013; Roter et al., 2002; Ferguson et al., 2002). Furthermore, studies from other industries indicate that women have more cautious approaches to solve complex problems (Powell and Ansic, 1997; Barber and Odean, 2001; Charness and Gneezy, 2012). So, we hypothesize that female physicians will order more diagnostic tests and distribute their daily load more evenly due to the effects discussed above. We develop two separate models in order to test our hypotheses related to number of diagnostic test orders and distribution of daily load.

We contribute to the literature on physicians' gender and experience effects on practice variations by using patient level retrospective data instead of survey data or physician level data which are commonly used in the literature; and by using operational variables such as finished and unfinished workload of the physician at the time of patient examination instead of using only physician, patient and hospital related control variables. Our aim is not to analyze whether physicians order adequate number of tests or unnecessary tests, nor to evaluate the impact of requested tests on the outcome of patients' complaints or quality of care. Because of increasing cost of diagnostic test orders, we would like to investigate whether physician characteristics such as experience and gender creates practice variations in terms of test ordering behavior; in particular we would like to identify characteristics that lead to greater number of diagnostic tests. This knowledge would be useful for healthcare managers in developing new strategies and policies for the use

of healthcare resources, as well as design and implementation of interventions to transform physician behavior (Hartley et al., 1987).

We used data from 5 outpatient units (internal medicine, gynecology, pediatrics, hematology, endocrinology) of a training and research hospital (see Section 4.4 for more details). Our motivation for choosing these outpatient units is the high use of diagnostic tests in decision making. Our study uses data for 2016 and includes 119,254 patient records in total after some exclusions to investigate whether there is association between diagnostic test ordering behavior of physicians and physicians' experience and gender. While we use patient level data in order to test our hypotheses related to number of diagnostic test orders, we use variables for each physician-day in order to test our hypotheses related to distribution of daily load.

We find evidence that there is association between diagnostic test ordering behavior of physicians and physicians' experience and gender. According to predicted results of our model, as physicians' experience, defined by years after graduation, increases from 0 to 35 years, mean diagnostic test orders increases from 8.5 to 10; and mean diagnostic test orders is 8.6 and 9.5 for male and female physicians respectively, while holding all the covariates at their means. We also find that physicians do not distribute their daily load evenly and distribution of daily load is right skewed, in other words, they rush to finish their load regardless of their gender and experience, but how much they rush changes with their experience and gender.

As a further analysis, we study the effect of interaction between gender of patient and physician; the effect of work location on test ordering behavior. We also conducted robustness checks by using alternative workload and experience variables. We also study the interaction of physician characteristics with workload.

The rest of the chapter is structured as follows. The next two sections review the related literature (Section 4.2) and introduce our hypotheses (Section 4.3). The details of our empirical setting are given in Section 4.4. We provide details of our data set, the variables used in our analysis and model specification in Section 4.5. Section 4.6 presents results, while Section 4.7 includes further analysis that has supporting evidences to practice variations from different perspectives. Finally,

Section 4.8 concludes the chapter.

4.2 Related Literature

In this study, we focus on practice variations, in terms of test ordering behavior of physicians and distribution of daily load due to their gender and experience at outpatient units of hospital. Although delivering high quality of care is significant for all physicians, variability in medical practice due to physicians' characteristics such as gender and experience has been an interesting problem in the literature. Physicians choose among alternative treatment options and update their beliefs about treatments based on the outcomes, therefore they continuously learn through experience and their beliefs about treatment-patient matches affect treatment choices (Gong, 2017). Learning by doing or experiential learning is a well-known mechanism in different industries including healthcare. Although experience is defined as the cumulative production history of an individual, which give an individual a chance to improve skills at performing established routines and practices (Reagans et al., 2005), the evidence is mixed in the literature. Choudhry et al. (2005) review the literature in empirical studies investigating the relation between clinical experience and quality of care. Among 62 studies they reviewed, 52% reported decreasing performance with increasing experience for all outcomes assessed. While their comprehensive review focuses on the studies analyzing patient outcomes, our study focuses on test ordering behavior. In the literature, there exist studies focusing on variation in outcome, use of resources, referrals, number of test orders, quality indicators etc.

The findings of Tsugawa et al. (2018), Tsugawa et al. (2017a), Tsugawa et al. (2017b) are particularly related to ours; since, like our study, they use patient level data. However, they focus on the impact of age and gender on the mortality and readmission rates of surgeons, general practitioners and hospitalist physicians, rather than test ordering behavior. Tsugawa et al. (2018) study whether patients' mortality differs according to the age and gender of surgeons. Although evidence is mixed in the literature, they find that patients' mortality decreases with the age of surgeons

among surgeons with high operative volume; however, they find no relation for surgeons with low and medium operative volume. On the other hand, they find no evidence that mortality differs between patients treated by female versus male surgeons, unlike other studies in the literature which suggest that female surgeons may have lower mortality rate than male surgeons (Wallis et al., 2017).

In another study, Tsugawa et al. (2017b) find that readmission rates do not differ with the age of hospitalist; but older hospitalists have higher mortality rates, while this effect disappeared among the hospitalists who treated high volumes of patients. In addition, mortality and readmission also differ between the patients treated by female and male physicians (Tsugawa et al., 2017a), where female physicians have lower rates for both indicators. On the other hand, Jerant A (2013) reports that the gender of physicians (family medicine and general internal medicine) is not associated with mortality or health care utilization (total health care expenditure, prescription/drug expenditures). However, the data is collected through patient surveys in this study, so self-report bias is a limitation for the study.

Studies in the literature that focus on test ordering behavior for a specific disease have found mixed results regarding the impact of physician age and gender. For instance, younger physicians are more likely to appropriately screen for hypertension (Aubin et al., 1994), but physicians with more than 15 years of experience are more likely to test for proteinuria of diabetes (Streja and Rabkin, 1999). Andersen and Urban (1997) find that women with a male physician to have almost two times greater risk for not having received a mammogram in the past two years and that differences in mammography screening associated with physician gender are not eliminated by controlling for patient characteristics. They argue this difference might be attributed to two reasons: female physicians order more mammography for their patients or female physicians' encouragement of mammography is more persuasive for the patients than male physicians. In their meta-analytic review, Roter et al. (2002) report that in primary care female physicians allocate more time for communication than male physicians, on the other hand, in gynecology male physicians have closer relationship with their patients than female physicians.

Franks et al. (2000) find that higher rates of laboratory referrals by community physicians were related with female physicians and increasing years of experience. Bugter Maessen et al. (1996) state that test ordering behavior of general practitioners has a significant positive relation to years of experience and working hours per week. However, Salloum and Franssen (1993) find that doctors who had been practicing for less than 10 years tended to order more tests and more expensive ones, and female physicians ordered more than male physicians. On the other hand, Leurquin et al. (1995) state that physician characteristics such as experience are not associated or slightly associated with use of blood tests except for gender of physician with a low correlation and Hartley et al. (1987) report a weak relation between experience and laboratory test requests among general practitioners. In another study, Epstein et al. (1984) find that patterns of test use among internists for hypertension patients did not appear to change with experience, as measured by years since graduation. Besides effect of experience, Vinker et al. (2007) find that female physicians order more tests than male physicians; higher number of patients is associated with more tests per patient; urban clinic physicians order more than rural clinic physicians. Similarly, de Gracia Gomis et al. (1999) report that although female physicians asked for a higher number of test than male physicians, more tests are ordered for male patients than for female patients. Studies also find that male and female physicians have differences in their practice patterns such as female physicians may be more likely to follow evidence-based practice and clinical guidelines (Baumhäkel et al., 2009; Berthold et al., 2008; Kim et al., 2005), perform better on standardized examination and have better communication and counseling skills (Jerant A, 2013; Roter et al., 2002; Ferguson et al., 2002).

Verstappen et al. (2004) focus on professional and context related factors that cause variation in test ordering behavior of physicians. They conducted a cross sectional analysis by collecting data through surveying 229 general practitioners. They found that individual involvement in developing clinical guidelines, working with a problem-oriented laboratory order form and working in group practices rather than working alone were related with lower volume of test orders. This shows us

that awareness about the use of diagnostic test orders, education style of physicians, updating previous knowledge with developing clinical guidelines and evidence could be some factors affecting use of diagnostic test orders or other clinical resources.

Although above studies mention practice variations due to physicians' gender and experience, we could not find any study analyzing relation between physician characteristics and distribution of daily load specifically. Since we find that number of diagnostic test orders per patient increases with increasing load of physicians due to rushing in Chapter 3 and we study the effect pf physician characteristics on diagnostic test orders in this chapter, investigating how physicians distribute their daily load, and whether physicians' gender and experience are associated with this distribution could give us significant information to observe the rush effect and to have a comprehensive understanding of the issues discussed in this thesis.

Our study is different from all the studies reviewed above, since we used operational control variables such as finished and unfinished load and time of examination besides physician and patient characteristics. Furthermore, we used patient level retrospective data from the electronic medical records of the hospital. We do not use survey data, which could have self-report bias, or do not use periodic averages such as daily average number of tests per visit of the physician, which may cause loss of some information.

Though there are studies using operational data, we use patient level data, while these studies use physician level data. For instance, Salloum and Franssen (1993) use average number of patients seen per day by a physician; Vinker et al. (2007) use age adjusted number of patients allocated to each physician as physician workload, since number of patient encounters is not known; and Leurquin et al. (1995) use number of encounters per week as indicator of workload. On the other hand, we have the data regarding the number of patients examined till the examination of the patient (finished workload), and number of patients waiting for the physician at the time of examination of the patient (unfinished workload), even we have daily number of encounters for the patient's physician as an additional control variable.

4.3 Hypothesis Development

In this section, we present hypotheses of our study related to practice variations in association with physicians' gender and years of experience, and relate them to the relevant literature in previous section.

First, the evidence and arguments presented in the literature about practice variation of physicians due to experience of physicians is mixed. On the other hand, experiential learning is a very well-known mechanism by which clinicians achieve improvements in patient outcomes with increased examination experience (Huesch, 2009). Reagans et al. (2005) state that studies in the learning by doing literature indicate that as individuals gain experience with the task, the number of errors they make decreases at a decreasing rate. However, medical advances occur frequently and treatment standards change with new clinical evidence, and the explicit knowledge that physicians possess over the years may easily become out of date. Therefore, it is also probable that physicians with more experience may be less likely to provide technically appropriate care (Choudhry et al., 2005) and to update themselves in accordance with clinical developments, since physicians are more likely to choose a treatment if they have used it on previous patients and with which they have had better outcomes on similar patients (Gong, 2017). Furthermore, education methodology and training practices also change over the years, and these changes create cohort effects (Tsugawa et al., 2017b).

Furthermore, in a review of empirical studies investigating the relation between clinical experience and quality of care, Choudhry et al. (2005) observe that, among 62 studies they reviewed, 52% reported decreasing performance with increasing experience for all outcomes assessed. In addition, Aubin et al. (1994) state that younger physicians are more likely to appropriately screen for hypertension, but Streja and Rabkin (1999) find that physicians with more than 15 years of experience are more likely to test for proteinuria of diabetes. Franks et al. (2000) find that higher rates of laboratory referrals by community physicians were related with increasing years of experience. Bugter Maessen et al. (1996) state that test ordering behavior of general

practitioners has a significant positive relation to years of experience. However, Salloum and Franssen (1993) find that doctors who had been practicing for less than 10 years tended to order more tests and Epstein et al. (1984) find that patterns of test use among internists for hypertension patients did not appear to change with experience, as measured by years since graduation. Also, Rosenbaum (2017) states that less experienced physicians may overuse medical resources in order to compensate their lack of experience.

Below, we provide our first two hypotheses, which states that physicians' years of experience in practice cause practice variation in terms of number of diagnostic test orders and distribution of daily load. However, the effect of physicians' experience on the number of diagnostic test orders and distribution of daily load could be in two ways due to the factors discussed above.

Hypothesis 1.a: *Number of diagnostic test orders per patient increases, while physicians' years of experience in practice increases.*

Hypothesis 1.b: *Number of diagnostic test orders per patient decreases, while physicians' years of experience in practice increases.*

Hypothesis 2.a: *Physicians distribute their daily load more evenly, while physicians' years of experience in practice increases.*

Hypothesis 2.b: *Physicians distribute their daily load less evenly, while physicians' years of experience in practice increases.*

Second, we study the relationship between the gender of the physician and practice variation of physicians in terms of the number of test orders and distribution of daily load. Studies found that male and female physicians have differences in their practice patterns such as female physicians may be more likely to follow evidence-based practice and clinical guidelines (Baumhäkel et al., 2009; Berthold et al., 2008; Kim et al., 2005), perform better on standardized examination and have better communication and counseling skills (Jerant A, 2013; Ferguson et al., 2002). Furthermore, studies from other industries indicate that women have more cautious approaches to solve complex problems (Powell and Ansic, 1997; Barber and Odean, 2001; Charness and Gneezy, 2012). Therefore, female physicians could order more

diagnostic test orders for their patients and distribute their daily load more evenly due to their cautiousness. Also, female physicians could be affected by the test demand of their patients more than their male colleagues or patients might be more demanding when they are examined by a female physician compared to a male physician, since female physicians have closer communication with their patients (Roter et al., 2002).

There are studies in the literature supporting our reasoning. Vinker et al. (2007), Salloum and Franssen (1993) and de Gracia Gomis et al. (1999) report that although female physicians asked for a higher number of test than male physicians. Andersen and Urban (1997) find that women with a male physician to have almost two times greater risk for not having received a mammogram in the past two years and that differences in mammography screening associated with physician gender are not eliminated by controlling for patient characteristics.

Below, we provide our hypotheses related to the effect of gender, which states that female physicians will order more diagnostic tests, due to the effects discussed above.

Hypothesis 3: *Female physicians order more diagnostic tests orders per patient compared to their male colleagues.*

Hypothesis 4: *Female physicians distribute their daily load more evenly compared to their male colleagues.*

4.4 Study Setting

The setting of this study is the outpatient units (polyclinic) of internal medicine, gynecology, pediatrics, hematology and endocrinology departments of a large training and research state hospital including its district polyclinics in the same city. The working hours at the outpatient units is from 8:30 to 17:30.

For the study setting, we selected the polyclinics, that order diagnostic tests most frequently. In this setting, there are 42 physicians consisting of 14 gynecology, 12 internal medicine, 2 endocrinology, 4 hematology and 10 pediatrics physicians. There are 11 male and 31 female physicians with on average 18 and 16 years of

Table 4.1: Information regarding physicians

	Physician Gender	
	Male	Female
Physicians (by departments)	11	31
Gynecology	3	11
Internal Medicine	5	7
Endocrinology	1	1
Hematology	2	2
Pediatrics		10
Av. Experience (in years)	18	16
Patients		
Male	26%	25%
Female	74%	75%
Age (except pediatrics)	43.1(43.1)	36.8(42.8)

experience after graduation respectively (Table 4.1). In our data set, of the 119,254 patient visits, 75% were female, while 25% were male, and they had an average age of approximately 38.6. The percentage of female patients is high, since 33,5% of the data belongs to gynecology department. We also observe that male-female distribution of patient visits for male (26-74%) and female (25-75%) physicians is very similar and close to general distribution of patients' gender, as it is shown in Table 4.1. Furthermore, 30% of male patient visits are to male physicians, while 70% of male patient visits are to female physicians. The ratio is very similar for female patient visits. 29% of female patient visits are to male physicians, while 71% of female patient visits are to female physicians. Average age of patients examined by male and female physicians is very close (43.1 and 42.8 respectively) when we do not consider pediatric patients, since all pediatric physicians are female. If it was different, our analysis could be biased. Figure 4.1 shows the number of patients examined in accordance with physicians' years of experience. However, we would like to note that number physicians at each experience level is not same. Table 4.2 presents the number of physicians at each experience level; how patients' age

changes with the experience of the physician, including pediatrics, in order to see whether patients' preference for the physician changes with their age. For the ease of representation, we divided physicians by their experience groups per 5 years from 0 to 35, though we use physician experience as discrete variable in the analysis. We observed no pattern in the distribution of patient's age in accordance with physicians' experience. We also note that our data is from 119,245 patient visits; hence, some patients come more than one time in a year and are counted multiple times.

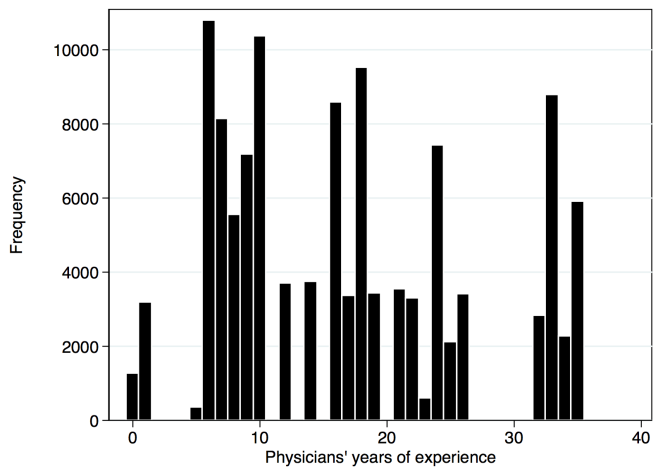


Figure 4.1: Histogram of number of patients examined at each physicians' years of experience

As mentioned before, the data belongs to the main hospital polyclinics and its district polyclinics. Table 4.3 shows the number of patient visits to each hospital by specialty. Some physicians work only at one location, while some of them work at multiple locations. Details of patient encounters per physician per specialty and per hospital are provided in Appendix E.1.

In this setting, the number of patients examined per physician in a given day ranges between 2 and 88 with a mean of 22.5 patients. We refer to the half day before noon as a day; hence, daily means before noon in this study as it is explained in the next section. Corresponding statistics per specialty and hospital are provided in Table 4.4.

Table 4.2: Summary of patient's age by physicians' experience (categorized)

Experience	# of Physicians	# of Patients Mean	Std. Dev.	Frequency
0 to 5	3	34.6	11.9	4,782
6 to 10	13	39.8	19.1	42,125
11 to 15	3	42.6	17.6	7,442
16 to 20	9	38.0	16.8	24,880
21 to 25	5	42.6	21.6	16,950
26 to 30	2	42.2	14.3	3,406
31 to 35	7	32.3	21.9	19,779
Total	42	38.6	19.4	119,254

Table 4.3: Number of patient visits per hospital

	Specialty					
Hospital	Endocrinology	Hematology	Internal Medicine	Gynecology	Pediatrics	Total
Main	1,998	9,998	21,644	31,909	11,197	76,746
District 1			11,383	3,110	603	15,096
District 2			12,910			12,910
District 3			4,243	3,807	1,882	9,932
District 4			3,305	1,265		4,570
Total	1,998	9,998	53,485	40,091	13,682	119,254

Table 4.4: Average daily load per physician

Specialty	All Hospitals	Main	District 1	District 2	District 3	District 4
Gynecology	29.1	37.8	16.1		15.9	11.7
Internal Medicine	28.9	30.4	29.9	36.9	19.7	17.6
Hematology	17.2	17.2				
Pediatrics	12.2	14.5	4.4		8.8	
Endocrinology	5.3	5.3				
Total	22.5	23.4	21.3	36.9	14.9	15.4

There exists previous literature which suggests physicians' test ordering behavior can be affected by profit concerns (Silverman and Skinner, 2004; Dafny, 2005; Powell et al., 2012). However, the hospital in our study is a public research hospital; and the physicians are civil servants, and have no financial incentives to advise for or against any particular course of treatment, including ordering more or fewer diagnostic test orders. The total number of patients that they can see a day is also determined by hospital management. Physicians are paid based on a fixed salary plus a share from the hospitals total revenues, and the diagnostic test orders have a negligible effect on their total earning. Hence, in our study, we can assume financial incentives do not have an effect on the test ordering decisions of physicians.

4.5 Methods

In this section, we represent methods of our analysis related to the practice variations in terms of number of diagnostic test orders and distribution of daily load. We developed two separate models in order to test our hypotheses related to test ordering behavior and distribution of daily load. While we conduct our analysis for test ordering behavior by using patient level data, we conduct our analysis related to distribution of daily load by using data for each physician day.

4.5.1 Model Development for Test Ordering Behavior

4.5.1.1 Data and Variables

The data for this study is from an outpatient unit and encompasses records of 119,254 patients visits from January 2016 to December 2016, which is generated by 42 physicians from different specialties, including internal medicine, gynecology, pediatrics, hematology and endocrinology, of the same hospital, including its district polyclinics in the same city. For each patient record, the data included information on (1) patient gender, (2) patient age, (3) the attending physician, (4) gender of the attending physician, (5) experience of the attending physician, (6) the ICD code, (7) the registration time, (8) the examination time, (9) the test ordering

time, (10) the number of diagnostic tests ordered, (10) the types of tests ordered, (11) the cost of diagnostic test orders and (12) patient's total number of visits to the hospital, (13) physician's specialty, (14) location (main hospital or district polyclinics). Throughout the chapter, we use the individual patient record (PR) as the unit of analysis.

In order to have this study sample, we excluded the PRs for the days where physicians examined patients together with their assistants, since in those cases, the PRs do not reflect the actual number of patients cared by a single physician during the day. We also excluded the PRs after the lunch break. Patients who were examined by a physician in the morning usually return to show their test results in the afternoon, and are not recorded in the system again. Hence, the number of PRs recorded in the afternoon does not reflect the real workload. In the remainder of the chapter, we refer to the half day before noon as a day; hence, daily means before noon in this study. Finally, we also excluded the instances where a physician examined only 1 patient a day. All the graphs and descriptive statistics in the chapter reflect only the patients remaining after the exclusion.

Our data had PRs with around 500 different ICD codes. The most frequently used ICD codes are provided in Appendix E.2. In addition, we deducted the ICD codes that were assigned less than 10 times during the period represented in the study. This leaves a sample of 119,254 PRs. Note that, all the patients excluded from the analysis sample are included in the estimation of the workload measures, since these patients are still examined by the physicians.

Next, we provide an overview of our independent, dependent and control variables.

Independent Variables

The independent variables for our analysis are physicians' gender (*physiciangender*) and experience (*physicianexperience*) that is the years in practice after graduation. Although physicians' experience are generally defined as the years in practice or age of physician in the literature reviewed in Section 4.2, patient volume of each physician during those years may not be same. So, physicians who have been in practice

for the same years, could have different levels of experience considering experiential learning/ learning by doing. However, years in practice cover "the cohort effect" which determines the practice style of physician (Tsugawa et al., 2017b). If we have the data for each physician related to cumulative patient volume for the years in practice, we could also use it as a proxy for physician experience in order to make a robustness check of our analysis.

Dependent Variables

We use the number of diagnostic tests ordered for a patient i by a physician ($ntests_i$) as our dependent variable. We expect that the service content per patient, i.e., the number of diagnostic test orders, may vary with physician gender and experience. The average number of diagnostic test orders per patient is 10.98 in our study sample; Figure 4.2 provides the histogram of the number of test orders.

Table 4.5: Average number of diagnostic test orders

Specialty	Mean	Std. Dev.	Min	Max
Gynecology	4.26	6.39	0	66
Internal Medicine	16.21	14.68	0	86
Hematology	13.96	13.71	0	159
Pediatrics	16.66	9.28	0	84
Endocrinology	20.65	22.18	0	137
All	10.98	13.27	0	159

Control Variables

In addition to the variables of interest, we include a number of control variables in this study, which can be broadly categorized into: (1) physician related controls: $specialty_i$ (pediatrics, hematology, endocrinology, internal medicine or gynecology), $hospital_i$ (work location of the physician), (2) patient related controls: $patientage_i$, $patientgender_i$, $status_i$ (control or new patient), ICD_i (since the type and number of test orders for different diagnoses can be quite varied), $nvisits_i$ (which indicate the patient's previous number of visits to the hospital, in order to control for the patient's history at the hospital), (3) operational controls: $waiting_i$ (the number of

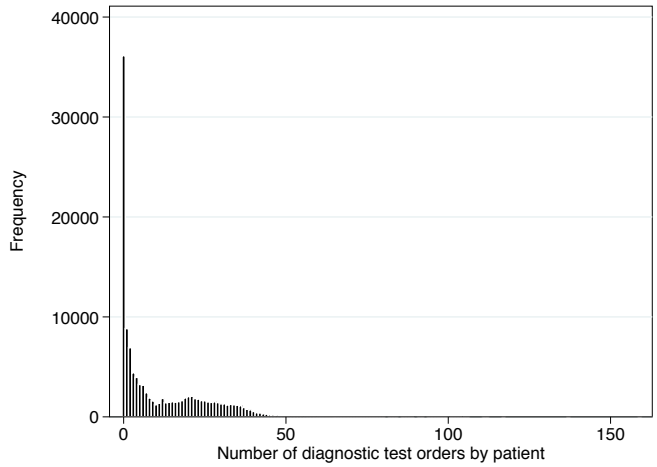


Figure 4.2: Histogram of number of diagnostic test orders per patient

patients in the waiting room while patient i is being examined by the physician), $finishedload_i$ (number of examined patients before patient i), $time_i$ (examination time interval of the patient i), and (4) time-related controls: $month_i$, day_i. While $specialty_i$, $hospital_i$, $status_i$, ICD_i, $patientgender_i$ are categorical variables, the rest of variables are numerical.

As argued in Chapter 3, operational variables, such as load ($waiting$) and finished load ($finishedload$) impact the number of tests ordered per patient; hence, unlike other studies in the related literature, which investigate the effects of physicians' gender and experience, we include these operational variables as controls in our model.

4.5.1.2 Model

We used negative binomial regression (NBREG) due to over-dispersion of the number of diagnostic test orders per patient, that is, the variance of number of test orders is much larger than its mean. In our study sample, we have 119,254 patient visits, which are generated by 42 physicians. The average number of patient encounters per physician is 2,842 and the maximum number of patient encounters per physician is 7,835. Besides controlling physicians' characteristics such as gender, experience,

specialty; in order to capture individual test ordering behavior of physicians, we also ran a negative binomial model with random effects (XTNBREG). If this model is significantly different than negative binomial model, then individual approach of physicians is also significant.

Instead of using zero inflated negative binomial (ZINB) model as in Chapter 3, we use NBREG in this chapter; since NBREG model provides random effects option and ZINB model is a highly complex model considering number of variables and size of data in this study. Furthermore, ZINB model in previous chapter does not converge, when we included all the ICDs (see Section 3.7.5); and the robustness checks in Section 3.7.1 show that similar results are obtained when NBREG model is used instead of ZINB model.

4.5.2 Model Development for Distribution of Daily Load

4.5.2.1 Arrival pattern of patients

First, we analyzed registration time, i.e., arrival pattern, of patients, since load perception of physicians could be affected by the arrival pattern of patients. Recall that physicians start working at 8:30 generally and take lunch break at 12:30. We used one hour time intervals, where 1 indicates 8:30-9:30 time interval. Figure 4.3 is the histogram for daily load of physicians, where the number of patients examined per physician in a given day ranges between 2 and 88 with a mean of 22.5 patients. As it is seen in Figure 4.4, patients start registering before physicians start working. Half of the patients complete their registration before 9:30, since median registration time is 1, while mean registration time is 1.2. Therefore, physicians start the day with a number of patients waiting for examination, which creates initial load perception of a physician. At time interval 1, physicians have 9 patients waiting on average. We analyzed whether patient arrivals are normally distributed for each specialty by using Kolmogrov-Smirnov test, Skewness-Kurtosis tests for normality as well as Shapiro-Wilk normality test. Results indicate that patient arrivals are not normally distributed in any of the specialties.

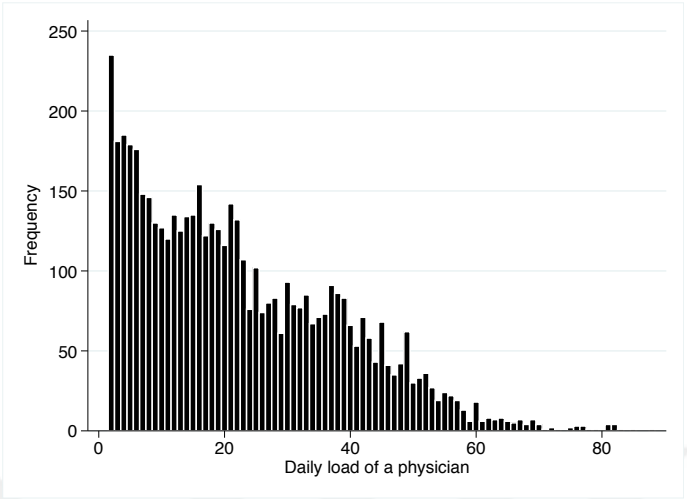


Figure 4.3: Histogram of daily load

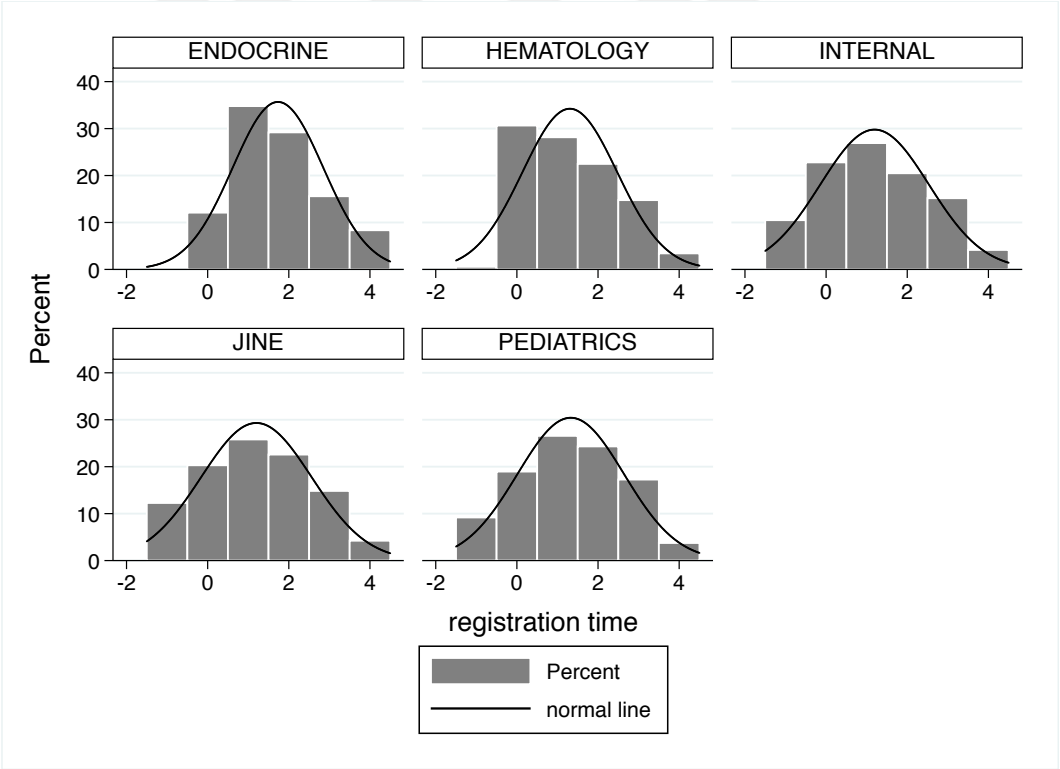


Figure 4.4: Histogram of registration times of patients by specialty

4.5.2.2 Distribution of load

Second, we analyzed examination time interval of patients in order to observe whether physicians distribute their load evenly from 8:30 to 12:30. We observe

that physicians may start working before 8:30. There are small number of patients (2%) examined before regular work hours, i.e., before 8:30. From Figure 4.5, we see that more than half of the patients are examined in the first two hours of day. Median examination time of patients is 2, while mean examination time of patients is 2.1.

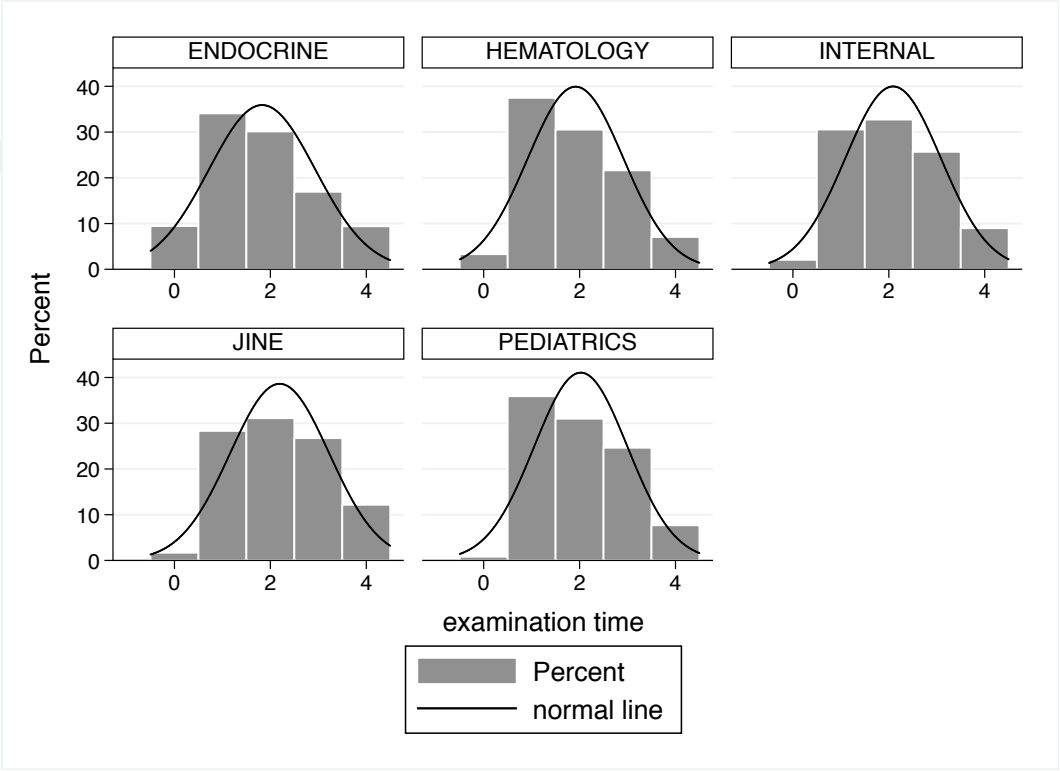


Figure 4.5: Histogram of examination times of patients by specialty

For further analysis, we draw quantile plots for each specialty in order to see whether patients are uniformly distributed over each time interval (Figure 4.6). We excluded patients examined before 8:30, since we do not want to distribute patients to this interval. In a quantile plot, each value of the variable is plotted against the fraction of the data that have values less than that fraction. The diagonal line is a reference line. If patients were rectangularly distributed, all the data would be plotted crossing the line. Because almost all the points are below the reference line, we know that the examination time of patients is right skewed, indicating that majority of patient are examined in the first half of the period. Since we observed

that patients are not uniformly distributed over each examination time interval, we also show that the data can be clearly distinguished from a uniformly distributed data by conducting Kolmogrov-Simirnov test for discrete uniform distribution.

4.5.2.3 Distribution of load at each physician day

After observing general distribution of load, we calculated skewness of patient distribution for every physician day by using SKEW formula in Excel . We called this as the skewness score of a physician day. If the skewness score is zero then the distribution of load is perfectly symmetric. If the skewness score is negative, then the distribution is left skewed, i.e., majority of patients are examined at the second half of the day; while if the skewness score is positive, then the distribution is right skewed, i.e., majority of patients are examined at the first half of the day. We have 4,983 physician days with mean skewness score of 0.316, where minimum skewness score is -2.828 and maximum skewness score is 3.606.

4.5.2.4 Model

In order to observe whether physicians' gender and experience affect skewness of the distribution of load, we conducted a random effects linear regression with the skewness score as dependent variable, physicians' gender and experience as independent variables and daily load, initial load (how many patients has already registered, when physician starts to work at 8:30), month, day of week, hospital and specialty as control variables while controlling for individual physicians as random effect. We used initial load as control variable, since the number of patients waiting in the beginning of the day could create initial sense of load and physicians may rush that results with increased skewness score.

4.6 Results

All of the estimations are done using STATA 14.1.

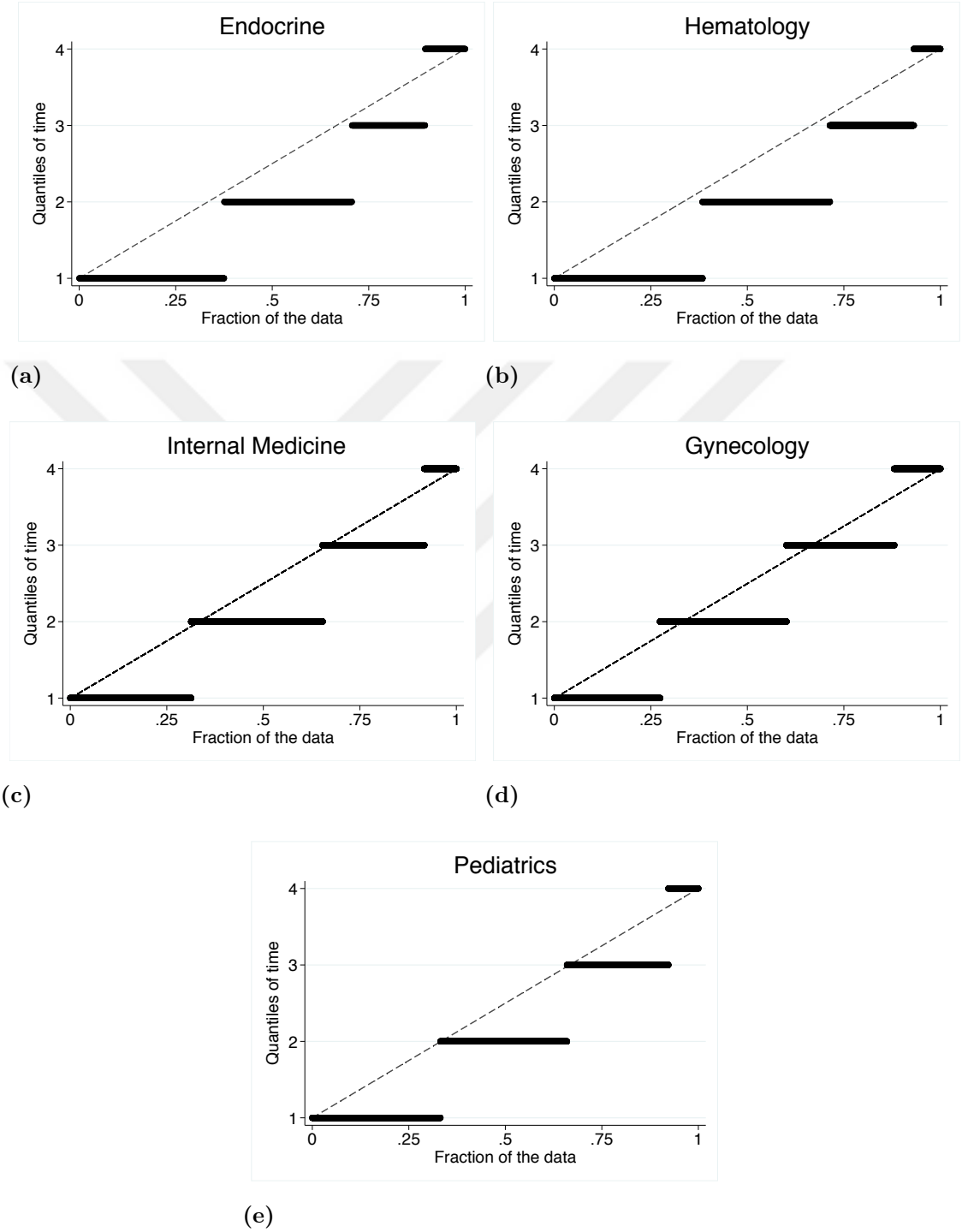


Figure 4.6: Quantile plots of examination times by specialty

4.6.1 Results of the Model for Test Ordering Behavior

The output of XTNBREG model includes a likelihood-ratio test, which compares the panel estimator with the pooled estimator (that is, a negative binomial estima-

tor). Accordign to likelihood-ratio test, XTNBREG model is significantly different than NBREG model, which shows that variability in test ordering behavior due to individual characteristic of physician is significant. Additionally, we compared AIC values of NBREG and XTNBREG models. Since NBREG model has higher AIC value than XTNBREG model, XTNBREG model is preferred over NBREG model. Therefore, we report only the results of XTNBREG (Table 4.6).

Table 4.6: Results of negative binomial model with random effects

Variables	XTNBREG
$physicianexperience$	1.005^{***}
	(0.001)
$physiciangender : female$	1.104^{***}
	(0.014)
$patientgender : female$	1.040^{***}
	(0.009)
$waiting$	1.005^{***}
	(0.001)
$finishedload$	0.997^{***}
	(0.000)
$time$	0.839^{***}
	(0.005)
$noofvisits$	0.999^{***}
	(0.000)
$status : new$	3.495^{***}
	(0.043)
Observations	$119,254$
# physicians	42
AIC	$679,053.2$
Log likelihood	$-339,269.59$
Standard errors in parentheses	
*** p<0.01, ** p<0.05, * p<0.1	

Table 4.6 provides incidence rate ratios (odds ratios) as the coefficients of variables. First, the significant coefficient of *physicianexperience* indicates that if a

physician's years of experience increases by one year, his rate for number of diagnostic test orders (*ntests*) would be expected to increase by a factor of 1.005, while holding all other variables in the model constant. This suggests our findings support Hypothesis 1a; that is, years of experience has a positive effect on the number of diagnostic tests ordered. Second, female physicians compared to males, while holding the other variable constant in the model, are expected to order 1.104 times more tests; thus supporting our Hypothesis 3. Additionally, to understand the results better, we can use the margins command to calculate predicted results at different experience levels and for each physician gender. We see that the mean number of diagnostic test orders for male physicians is 8.6 and for female physicians 9.5, while all other variables are at their mean values, i.e., *waiting* equals to 6.150, $finishedload$ equals to 16.592 etc.. We also obtain the mean number of diagnostic test orders for *physicianexperience* that range from 0 to 35, while all other variables are at their mean values. Mean number of diagnostic test orders increases from 8.5 to 10, when physicians' experience increases from 0 to 35 years.

We also would like to discuss the coefficients of our control variables briefly. Regarding the operational variables, number of diagnostic tests orders increases with increasing number of waiting patients (*waiting*), decreases with increasing number of examined patients ($finishedload$) and with increasing time interval of the day. These results confirm our findings in Chapter 3. Furthermore, as it is expected, number of diagnostic test orders per patient decreases with increasing number of patient visits, which suggests that previous records of the patient at the hospital is significant for the diagnostic decision process of physicians; and new patients are given 3.495 more diagnostic tests compared to control patients. Details about other control variables are represented in Appendix E.3.

We conclude this section with a brief discussion about multiple comparison tests for corrected p-values. Since we perform a large number of statistical tests in this analysis, some of the variables could have significant coefficients purely by chance. Table 4.6 represents the results according to uncorrected p-values. Since Stata does not have command to perform multiple comparison tests, we conducted Bonferroni

correction to obtain the Bonferroni adjusted p-values where we multiply the uncorrected p-values with the total number of variables (around 500) in the model by ourselves. We observed that significance of variables represented in the Table 4.6 does not change after obtaining corrected p-values. For instance, *physicianexperience* and *physiciangender* have p-values of $4.44089e-16$ and $1.28131e-06$ respectively, when we multiply this with 500 we still have p value less than 0.01. Therefore, our results preserved after multiple comparison correction.

4.6.2 Results of the Model for Distribution of Daily Load

Results of the model indicate that female physicians have less skewed distribution of their daily load compared to their male colleagues, i.e., female physicians distribute their load more evenly over the day; and experienced physicians have less skewed distribution, i.e., with increasing experience physicians distribute their load more evenly (Table 4.7). Then, our findings support Hypothesis 4 and Hypothesis 2.a. However, when we calculated marginal effects of physicians' gender and experience at means of all other control variables, we observed load is right skewed at every experience level and for each physician gender (Appendix E.6), though skewness changes. We could claim that physicians rush to finish their load regardless of their gender and experience, but how much they rush changes with their experience and gender. We also observe that skewness score increases with increase initial load of a physician, i.e., physicians tend to rush more if they start day with higher number of registered patients. On the other hand, with increasing daily load skewness score decreases, i.e., physicians distribute their load more evenly with increasing daily load.

Our findings provide significant information for future research, which will be discussed in the next chapter (see Section 4.8).

We also conducted same analysis for the days with load more than 10, 15 and 20. We observed that while direction of coefficients for physicians' gender and experience does not change, their significance changes (Appendix E.7).

Table 4.7: Results of skewness model for distribution of daily load

Variables	Skewness Model
$physiciangender : female$	-0.062^{*}
	(0.036)
$physicianexperience$	-0.003^{**}
	(0.002)
$initialload$	0.019^{***}
	(0.003)
$dailyload$	-0.007^{***}
	(0.001)
Observations	4,983
R-squared	0.036
# physicians	42
Standard errors in parentheses	
*** p<0.01, ** p<0.05, * p<0.1	

4.7 Further Analysis

In this section, we provide further analysis of this study including the analysis for the effect of interaction between gender of patient and physician (Section 4.7.1) and work location on test ordering behavior (Section 4.7.2). We also conducted two robustness checks: instead of using continuous *physicianexperience* variable, we categorized physicians according to their years of experience as senior and junior physicians (Section 4.7.3); and instead of using two operational variables *waiting* and $finishedload$, we used daily load (*dailyload*) as the operational control variable (Section 4.7.4). We also analyzed whether recent experience (last 6 month) of physicians, instead of years of experience in practice, is significant (Section 4.7.5). We conducted another robustness check for the analysis related to distribution of daily load (Section 4.7.6). Additionally, interaction between physician characteristics and workload is analyzed (Section 4.7.7).

4.7.1 Interaction Model for Physician-Patient Gender

Additionally, we also checked whether there is any interaction between gender of patient and physician which affects test ordering behavior (Table 4.8). While holding all the variables same in the main model (XTNBREG), we added an interaction term for gender of physician and patient.

The results presented in Table 4.8 shows interaction between two binary variables, i.e., male-female physicians and male-female patients. Female physicians compared to male physicians are expected to order at a rate of 1.093 times greater number of diagnostic tests for male patients, i.e., (*physiciangender*: female) - (*physiciangender*:male) for (*patientgender*: male) = 1.093. Also, male physicians order 1.031 times greater number of diagnostic test orders for female patients compared to male patients, i.e., (*patientgender*: female) - (*patientgender*:male) for (*physiciangender*: male) = 1.031. However, the interaction term is not statistically significant, i.e., female physicians compared to male physicians order 1.013 times greater number of diagnostic test for their male and female patients but it is statistically significant ([(*physiciangender*: female) - (*physiciangender*:male) for (*patientgender*: female)]-[(*physiciangender*: female) - (*physiciangender*:male) for (*patientgender*: male)] = 1.013). Our main model also has the same results, i.e., female physicians order more and female patients have more diagnostic test orders. Additionally, Akaike's information criteria and log likelihood of the model is not better than our main model.

4.7.2 Practice Variations at Different Work Locations

Our findings in the main model showed that hospital code is a significant control variable. Then, we decided to make further analysis in order to see whether same physicians have practices variations at different locations, since patient profiles, access to laboratories, both technical and managerial system differences may possibly cause practice variation. We observed that there are 10 pediatric physicians working at 3 different locations. However, when we run our model only for these physicians, we observed no practice variations due to the work location in this setting. In order

Table 4.8: Gender interaction model

Variables	Interaction Model
physicianexperience	1.005***
	(0.001)
physiciangender : *female*	1.093***
	(0.020)
patientgender : *female*	1.031**
	(0.015)
patientgender : *female*#*physiciangender* : *female*	1.013
	(0.017)
Observations	119,254
# physicians	42
AIC	679,054.6
Log likelihood	−339,269.31
Standard errors in parentheses	
*** p<0.01, ** p<0.05, * p<0.1	

to increase number of physicians, we added 12 gynecology physicians working at the same locations. When we run the model with these 22 physicians, we again observed no practice variations between these 3 work locations (Appendix E.4). Since we made the analysis with limited number of physicians, as a future research we could study practice variations of physicians due to work location with a larger data set. Furthermore, these 3 work locations may not have significantly different work environments.

4.7.3 Comparison of Senior and Junior Physicians

According to our main model, as physicians' years of experience increases, number of diagnostic test orders per patient increases. We support our findings by comparing two groups of physicians. First group includes 16 physicians with less than 10 years of experience and second group includes 14 physicians with more than 20 years of experience in practice. We ran our XTNBREG model by using categorical experience variable for these two group of physicians instead of using continuous

experience variable (*physicianexperience*). The results indicate that second group of physicians (senior) order 1.123 times more diagnostic tests orders per patient compared to first group (junior).

4.7.4 Alternative Load Variable

We confirmed our results with alternative load variable, instead of using $waiting_i$ and $finishedload_i$, which are the instant values related to unfinished and finished load of the physician when examining patient i, we used daily load ($dailyload_i$) of the physician at the examination day of the patient i. Table 4.9 represents the results of the alternative model, which confirms our results, presented in Table 4.6, i.e., female physicians order more than male physicians; number of diagnostic test orders increases with the experience of physician. Additionally, the model shows that increasing daily load results with increased number of diagnostic test orders per patient.

Table 4.9: Alternative model with daily load of the physician

Variables	Alt. Load Model
physicianexperience	1.005***
	(0.001)
$physiciangender : female$	1.108***
	(0.014)
$patientgender : female$	1.039***
	(0.009)
dailyload	1.001***
	(0.000)
Observations	119,254
# physicians	42
AIC	679,158.3
Log likelihood	−339,323.15
Standard errors in parentheses	
*** p<0.01, ** p<0.05, * p<0.1	

4.7.5 Alternative Experience Variable

Huckman and Pisano (2006) highlight that there are studies in the literature that uses volume of patients cared in recent year as measure of experience of a physician rather than cumulative volume, because there are studies suggesting that experience may depreciate over time. Huesch (2009) also states that forgetting is also another effect that causes depreciation of experience. Therefore, recent volume of patients could be a predictor of experience as well. By using our data set, we calculated the experience as the number of patients cared in the last 6 month. However, our data set diminishes to around 35,000 PRs of 23 physicians, since PRs in first 6 month could not be used and some physicians do not work full time for the last 6 month. When we run the model with the new experience variable $experience6month$, instead of $physicianexperience$, we observe that $experience6month$ has almost no effect on the number of diagnostic test orders with the coefficient of 1.000 although it is significant, while the direction of other variables are preserved (See Appendix E.5). Then, we conclude that our findings contradict with claims of Huckman and Pisano (2006) and Huesch (2009). Recent experience of physicians does not affect ordering behavior of physicians notably. However, this analysis is done with a limited data set. Then, as a future study, this analysis could be repeated with a larger data set. Also, instead of using experience in the last 6 month, we could use the number of patients examined in the last one or two year right before the patient examined in the data set as the experience level of a physician.

4.7.6 Robustness of Skewness Score Model

As a robustness of skewness analysis, we also used chi-squared goodness of fit test as an approach to observe how much each physician day is close to uniform distribution. We calculated chi-squared statistic for each physician day[1]. In this analysis, we used

[1]In calculation of chi-squared statistic, square of the deviation from uniform distribution of daily load for each time interval is summed after dividing by the expected frequency to weight frequencies. If the sum is smaller than critical value of chi-squared distribution, there would be no reason to reject that distribution of daily load is uniform.

chi-squared statistic of each physician day as a proxy for how physicians distribute their daily load. Instead of skewness value in previous regression analysis, we used chi-squared statistic. We run the same model for the days with load more than 10, 15 and 20 as well. Results indicate that female and male physicians do not have significantly different daily load distribution, however experienced physicians have smaller chi-square value, i.e., experienced physicians distribute their daily load more uniformly (Table 4.10). Also, with increasing initial load and daily load, chi-squared value increases, i.e., distribution of daily load becomes more different than uniform distribution (Appedix E.8).

Table 4.10: Results of chi-squared model for distribution of daily load

Variables	Chi-Square Model
$physiciangender : female$	−0.071
	(0.416)
$physicianexperience$	−0.046***
	(0.018)
$initialload$	0.222***
	(0.020)
$dailyload$	0.024***
	(0.009)
Observations	5,339
R-squared	0.202
# physicians	42
Standard errors in parentheses	
*** p<0.01, ** p<0.05, * p<0.1	

4.7.7 Interaction of Physician Characteristics with Workload

According to the results of test ordering behavior model and distribution of daily load model, female physicians and experienced physicians order more diagnostic test orders compared to male physicians and less experienced physicians respectively; and female physicians and experienced physicians have less skewed distribution of

load, i.e., they rush less. In Chapter 3, we conclude that with increasing workload physicians order more diagnostic test orders due to rushing, since physicians use diagnostic tests as a substitute for time with the patient. According to these findings we expect to observe that if a physician rushes less, he/she orders less. However, we observe that even female physicians and experienced physicians rush less compared to male physicians and less experienced physicians respectively, they order more. Hence, we consider that gender and experience are significant physician characteristics affecting test ordering behavior. Then, we conduct another analysis regarding the interaction between physician characteristics and workload. We run our test ordering behavior model after adding interaction terms for *physiciangender* and *waiting* as well as *physicianexperience* and *waiting*. The results of the model is presented in Table 4.11. While the effects of all variables are preserved, we observe that there are significant interaction between workload and experience as well as female physicians. However, we would like to note that their effects are quite minor. The effect of one year increase in years of experience would be expected to change by a factor of 1.000 (0.99979) for each increase in the number of waiting patients and vice versa. Although it is significant, interaction has almost no effect. On the other hand, the effect of being female physician would be expected to increase by a factor of 1.003 for each increase in the number of waiting patients compared to being male physician and vice versa. Furthermore, marginal effect of physicians' gender and experience are positive and significant.

4.8 Discussion and Conclusion

This study contributes to medical literature regarding how physicians' characteristics such as experience and gender affect physician practices in terms of diagnostic test ordering behavior and distribution of daily load while taking physicians' workload into consideration. In particular, using data from the outpatient units of a large training and research hospital, we find that, the number of diagnostic tests ordered per patient and distribution of daily load, change in response to the physicians' experience and gender. Although it is difficult to make strong conclusions about the

Table 4.11: Results of model for test ordering behavior with interactions

Variables	XTNBREG with Interactions
physicianexperience	1.006***
	(0.001)
physiciangender : *female*	1.032***
	(0.015)
physicianexperience#waiting	1.000***
	(0.000)
physiciangender : *female#waiting*	1.003***
	(0.001)
waiting	1.008***
	(0.001)
finishedload	0.996***
	(0.000)
time	0.883***
	(0.004)
patientgender : *female*	1.037***
	(0.008)
Observations	119,254
# physicians	42
AIC	737,419.8
Log likelihood	−368,450.89
Standard errors in parentheses	
*** p<0.01, ** p<0.05, * p<0.1	

differences in test ordering behavior of physicians in accordance with the physicians' gender and experience, our results suggest that the difference in practice style in terms of use of diagnostic test orders may affect quality of care, patient outcomes, patient satisfaction and cost of care in accordance with the related literature.

Our findings may have many possible explanations. First, we found that less experienced physicians order fewer tests. Difference in physician practices in accordance with physicians' experience may occur due to changing training practices, education methods and treatment standards over the years as well as due to learning by doing/experiential learning that is found in many industries. Furthermore, Salloum and Franssen (1993) claim that if a group of physicians is trained to practice defensive medicine, they will continue to practice same way no matter what the real risk of malpractice and litigation. Therefore, we could conclude that education style of a physician significantly affects the practice style and physicians, educated around the same years, are expected to practice in a similar way, which is called as "age effects" and "cohort effects". In addition, experienced physicians could be more suspicious due to their previous experiences and could order more.

Second, taken together with previous evidence (Tsugawa et al., 2017a), our findings indicate that physicians' gender is also a significant determinant of practice style. In addition to studies from medical literature (Baumhäkel et al., 2009; Ferguson et al., 2002; Bertakis et al., 1995; Roter et al., 2002; Roter and Hall, 2004), studies from other industries also suggest that women have more cautious approaches to solve complex problems (Powell and Ansic, 1997; Barber and Odean, 2001; Charness and Gneezy, 2012). If these findings also apply to how male and female physicians approach to clinical problems and decisions in outpatient unit, such behavior may provide a mechanism linking why female physicians order more diagnostic tests compared to male physicians.

We also find that practice variation in terms of distribution of daily load occur in response to physicians' gender and experience. Female physicians distribute their daily load more evenly compared to male physicians, and experienced physicians distribute their daily load more evenly compared to less experienced ones, though

all physicians regardless of their gender and experienced have a right skewed distribution of load. These findings provide significant information. In Chapter 3, we conclude that with increasing workload physicians order more diagnostic test orders due to rushing, since physicians use diagnostic tests as a substitute for time with the patient. In Chapter 4, we find that female physicians and experienced physicians order more diagnostic test orders compared to male physicians and less experienced physicians respectively; and female physicians and experienced physicians have less skewed distribution of load, i.e., they rush less. According to our these findings we expect to observe that if a physician rushes less, he/she orders less. However, we observe that even female physicians and experienced physicians rush less compared to male physicians and less experienced physicians respectively, they order more. Hence, we consider that gender and experience are significant physician characteristics affecting test ordering behavior. Actually, these findings are in line with the literature, since females have more cautious approaches to solve complex problems (Powell and Ansic, 1997; Barber and Odean, 2001; Charness and Gneezy, 2012), they could rush less and order more compared to males. As a future study, we could analyze female physicians and experienced physicians among themselves, where we expect to observe female physicians/experienced physicians with higher skewness score order more diagnostic test orders compared to female physicians/experienced physicians with lower skewness score.

Although it is hard to eliminate differences in physicians' practices, we could have some suggestions to decrease unnecessary use of diagnostic test orders in accordance with the previous literature. Verstappen et al. (2004) state that social influence of colleagues among general practitioners is a significant determinant of test ordering. Therefore, increasing collaboration between male and female physicians as well as between senior and junior physicians could decrease the differences in practice style. In addition, interventions such as providing feedback on test ordering, starting a quality improvement program regarding use of medical resources such as diagnostic test orders, developing guidelines in accordance with up to date evidence-based medicine for the use of diagnostic tests, incorporating all the physicians while developing

these programs and guidelines and organizing follow up meetings about the results of feedback reports, newly applied programs and guidelines appear to have promising results to decrease unnecessary or overuse of diagnostic tests. Bugter Maessen et al. (1996) also suggest that feedback is expected to improve the rationality behind test ordering behavior and to reduce the number of unnecessary orders. Although these interventions could affect physicians in a different way, we expect to increase physicians' sensitivity about increasing use of medical resources and cost of care.

Finally, we would like to mention the limitations of our study. Although we made our analysis by using around 119,000 patient visits, the data is generated by 42 physicians who are from 5 different specialties. This study could be repeated by increasing number of physicians and focusing only one specialty. As a future research, we could expand our data set with additional physicians from other hospitals.

Chapter 5

CONCLUSION

This thesis investigates the mechanisms behind practice variation of physicians in terms of use of diagnostic test order, since increasing healthcare expenditures is one of the mostly discussed issues for the future of healthcare and over use of diagnostic test orders is reported in the literature (Zhi et al., 2013; Miyakis et al., 2006; Fryer and Smellie, 2013; Rosenbaum, 2017). Over-testing is undesirable not only because of its financial cost, but also because of its cost to the system and patients such as the anxiety due to false-positive and false-negative results, wrong interventions, undervaluation of clinical examination, overcrowding of laboratories, delays in diagnosis (Miyakis et al., 2006). Hence, analyzing test ordering behavior of physicians deserves scientific inquiry. This thesis empirically investigates how diagnostic test ordering behavior of physicians changes in the context of outpatient unit in response to workload and physician characteristics such as gender and years of experience in practice. It is significant to determine the mechanisms behind test ordering behavior in order to provide information to decision makers while developing new strategies and policies for the use of healthcare resources, as well as design and implementation of interventions to change test ordering behavior of physicians and to reduce healthcare costs. While suggesting solutions, we need to know that there are many mechanisms shaping physicians test ordering behavior. Knowing this fact is significant to avoid oversimplified solutions (Rosenbaum, 2017).

In Chapter 3, we study how workload, which is divided into two as finished and unfinished load, affect test ordering behavior of physicians and in Chapter 4, we study how physicians' gender and experience affect test ordering behavior and distribution of daily load. The study presented in Chapter 3 contributes to the growing body of literature that suggests workers are not state independent and

show adaptive behavior in response to changes in the system, such as the amount of workload (Schultz et al., 1998, 1999; Freeman et al., 2017; Batt and Terwiesch, 2016; Berry-Jaeker and Tucker, 2017). In particular, there are also studies in the context of diagnostic test ordering behavior of clinicians (Deo and Jain, 2018; Alizamir et al., 2013). Our study differs from these studies by suggesting diagnostic test orders as a substitute for time with the patient, and investigating the effect of workload and finished load on the content of work, rather than service time. Furthermore, we use finished load as a proxy for fatigue in this empirical study, while studies of fatigue in prior medical literature rely on self-reported measures (Mazur et al., 2016; Gaba and Howard, 2002). Additionally, the study in Chapter 4 contributes to medical literature regarding how physicians' characteristics such as experience and gender affect diagnostic test ordering behavior and distribution of daily load while taking physicians' operational variables into consideration, which is missing in the previous literature. This study also uses patient level retrospective data, while the previous literature mostly depends on surveys and self-reported measures.

In this chapter, we conclude this thesis by synthesizing our empirical findings with the possible factors under these findings, talking about the implications and limitations of the study and direction of future research.

In Chapter 3, using data from the internal medicine unit of a public hospital, we find that, the content of the work, that is, the number of diagnostic tests ordered per patient, changes in response to workload. In particular, we find that the number of diagnostic test orders increases with workload, measured by the number of patients waiting to be examined. This might be due to various concerns of physicians, such as the need to serve all waiting patients in a limited time, a desire to improve diagnostic accuracy under time constraints and increasing patient satisfaction/perception of quality of service. We also find that, physicians order fewer tests per patient as the finished load, measured by the number of patients already examined increases. Fatigue and mental depletion may dominate concerns like malpractice, diagnostic accuracy, and patient satisfaction, leading physicians to be less engaged with patients, and to start ordering fewer tests.

In addition, our findings in Chapter 4 indicate that the number of diagnostic tests ordered per patient, changes in response to the physicians' experience and gender. Using data from five outpatient units (internal medicine, gynecology, endocrinology, hematology, pediatrics) of a public hospital, we find that female physicians and more experienced physicians order more diagnostic tests per patient. We also found that female physicians and experienced physicians distribute their daily load more evenly. Difference in physician practices in accordance with physicians' experience may occur due to changing training practices, education methods and treatment standards over the years as well as due to learning by doing/experiential learning that is also found in many industries. Regarding the practice variation due to physicians' gender, studies from other industries suggest that women have more cautious approaches to solve complex problems (Powell and Ansic, 1997; Barber and Odean, 2001; Charness and Gneezy, 2012) and there are studies in medical literature supporting our findings and indicating that female physicians are more likely to follow clinical guidelines and evidence-based practice (Baumhäkel et al., 2009; Berthold et al., 2008; Kim et al., 2005), perform better on standardized examination and have better communication and counseling skills (Jerant A, 2013; Roter et al., 2002; Ferguson et al., 2002). All these factors are possible reasons for test ordering behavior of female physicians.

Empirical findings of this thesis show that in order to develop effective strategies and policies to change test ordering behavior of physicians we need to know mechanisms behind it. In accordance with our findings, we suggest some managerial and policy implications. First, workload effect may be reduced by interventions to increase available physician time per patient, such as reducing the number of patients allocated to a given physician, or using appointment systems, which would decrease the number waiting in the clinic, thus alleviating the rush effect. Recall that, in accordance with the results of the empirical model studied in Chapter 3, we show that using appointment system, instead of first come first served system in the outpatient unit, has the potential to decrease number of diagnostic tests by 1.3 per patient (see Section 3.8 for details). Considering total number of patients cared in the polyclinic for whole day (around 100 patients by 3 physicians), the

hospital has the opportunity to save the cost of around 130 diagnostic tests per day if they switch to appointment system fully. Global effect of this change seems to be non-negligible, so it deserves attention for further research. Second, although it is hard to eliminate differences in physicians' practices due to physicians' gender and experience, increasing collaboration between male and female physicians as well as between senior and junior physicians could decrease the differences in practice style. Verstappen et al. (2004) also state that social influence of colleagues among general practitioners is a significant determinant of test ordering. For additional suggestions and recommendations see Section 3.8 and Section 4.8.

Our findings provide insights for recent discussions about rational and effective use of medical resources as well as about clinical prevention. In order to improve the rationality behind use of diagnostic tests, interventions such as defining paths for diagnostic decisions, providing feedback on test ordering, defining key performance indicators related to use of diagnostic tests, and sharing regular reports and statistics with physicians about these indicators could increase physicians' awareness about increasing use of medical resources and cost of care. Additionally, clinical prevention is evaluated in three levels until today: primary prevention, secondary prevention, tertiary prevention. In recent years, quaternary prevention has been raised and in its definition, it includes actions taken to identify patients at risk of overmedicalization (Martins et al., 2018). Therefore, we could evaluate over-testing from this perspective as well. If policy makers or healthcare managers develop strategies regarding quaternary prevention actions, physicians keep these actions in their minds for every intervention they suggest to a patient. As Martins et al. (2018) mention patients may suffer harm from medical interventions, so through quaternary prevention, we could avoid over-testing and undesired harm to the patient due to over-testing.

On the other hand, Rosenbaum (2017) states how hard eliminating the waste safely is, because interventions could take us to oversimplification which might cost more than overuse. While she mentions a theory that "if high spenders would just behave like low spenders, we could save $700 billion a year without compromising quality", it is significant to note that we cannot account all the variation as waste.

Quality and outcome based studies will enhance the findings of this study in order to suggest efficient policies. While we study the effect of workload and physician characteristics on test ordering behavior, there are numerous factors affecting physicians.

The studies in this thesis have some limitations, though we tried to overcome some of them by using control variables and conducting robustness checks. First, as we mentioned in Section 3.8, we are not able to control for the comorbidities of patients due to our data structure. However, comorbidities are not expected to be correlated with our independent variables ($waiting$, $finishedload$); hence, the coefficients of our covariates would not be effected by this omission. Since we control for age, ICD code and the number of visits to the hospital, which are potentially correlated with comorbidity of a patient, we partially control for comorbidity through these variables. However, we note that, the results related to these variables should be interpreted in the light of the possible bias. Another limitation of this study is the number of physicians that generates the study sample of 11,271 PRs of 6 physicians. Although we confirmed our findings using 55,424 PRs of 14 physicians from gynecology polyclinic, these findings should be confirmed in other settings with larger data sets. Second, as it is mentioned in Section 4.8, although our study sample includes around 119,000 patient visits, the data is generated by 42 physicians who are from 5 different specialties. As a future research, this study could be repeated by increasing number of physicians and focusing only one specialty.

In accordance with our findings and insights we gained from these studies, we come up with a future research direction. First of all, the effect of workload as well as the effect of gender and experience on physicians diagnostic test ordering behavior can also be studied through experimental studies, that give us opportunity to measure more variables and control the environment. For instance, in this data structure, we do not know exactly the length of examination for each patient which gives a chance for further analysis supporting our findings. In experimental study setting, we can design in accordance with the research question in order to collect the required data instead being limited with the retrospective data. In another study,

non-linear effect of workload could be studied. Also, we could use different measures for finished workload. For instance, instead of using number of examined patients simply as finished workload, we could add diagnostic complexity in accordance with the ICD code as a weight to each patient. In this case, we assume that complex cases will create more fatigue. Second, we could conduct studies related to cost of diagnostic test orders. For instance, whether workload affect choice of diagnostic test orders in association with their cost is a significant question to understand whether physicians increase the number of diagnostic test orders by ordering cheaper tests under high workload. A new study can be designed by focusing on a specific disease such as hypertension or diabetes and an analysis for each test requested for this disease can be made. Whether a test is ordered more at times when workload is high can be studied by developing a logit model while controlling for patient specific and operational variables. Also, in another analysis, we could use cost of diagnostic tests as dependent variable of the main model in Chapter 3. Additionally, this study could be repeated in different work environments such as private hospitals where physicians could have different motivations and hospitals where physicians have daily quota for the total cost of diagnostic test ordered. Third, another study, which we could not make due to limitations in our data, could be conducted to compare diagnostic decisions of physicians before and after the lunch break.

Another research question might be related to the variance in physicians diagnostic test ordering behavior for a specific disease under workload. If the variance is low under high workload condition and the variance is high under low workload condition, we could consider that physicians are able to make patient specific decisions when they have time to communicate with the patient. Fourth, although we are not focusing on patient outcomes, patient satisfaction, physician satisfaction or quality of care in this thesis, the practice variations due to workload could possibly affect these. These issues should be studied as a future research in order to highlight the results of these practice variations.

In summary, under the discussions about increasing healthcare expenditures and increasing share of healthcare spending in countries GDP, it is important to analyze

the possible factors behind it in order to suggest effective strategies and policies to control healthcare expenditures or even to grab healthcare managers and decision makers' attention on the issue. In this thesis, empirical investigation of how workload and physician characteristics affect diagnostic test ordering behavior of physicians shows that increasing workload results in increasing number of diagnostic test orders per patient; increasing finished load results in decreasing number of diagnostic test orders per patient; and female physicians and experienced physicians order more diagnostic tests compared to male physicians and less experienced physicians respectively. Therefore, our findings contribute to the body of knowledge in both healthcare operations management and medical literature with showing state dependency of work content while considering diagnostic tests as a substitute for time and showing that characteristic of workers such as gender and experience are also covariates of work content.

BIBLIOGRAPHY

Alizamir, S., de V ericourt, F., and Sun, P. (2013). Diagnostic accuracy under congestion. *Management Science*, 59(1):157–171.

Andersen, M. and Urban, N. (1997). Physician gender and screening: do patient differences account for differences in mammography use? *Women Health*, 26(1):29–39.

Aubin, M., Ve'zina, L., Fortin, J., and Bernard, P. (1994). Effectiveness of a program to improve hypertension screening in primary care. *Canadian Medical Association Journal*, 150(4):509–515.

Barber, B. and Odean, T. (2001). Boys will be boys: gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, 116(1):261–292.

Batt, R. and Terwiesch, C. (2016). Early task initiation and other load-adaptive mechanisms in the emergency department. *Management Science*, 63(11):3531–3551.

Baumhäkel, M., Müller, U., and Böhm, M. (2009). Influence of gender of physicians and patients on guideline-recommended treatment of chronic heart failure in a cross-sectional study. *European Journal of Heart Failure*, 11(3):299–303.

Berry-Jaeker, J. and Tucker, A. (2012). Hurry up and wait: Differential impacts of congestion, bottleneck pressure, and predictability on patient length of stay. *Harvard Business School Working Paper,No. 13-052.*

Berry-Jaeker, J. and Tucker, A. (2017). Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Science*, 63(4):2392–2397.

Berry-Jaeker, J., Tucker, A., and Lee, M. (2013). Increased speed equals increased wait: The impact of a reduction in emergency department ultrasound order processing time. *Harvard Business School Working Paper,No. 14-033*.

Bertakis, K., Helms, L., Callahan, E., Azari, R., and Robbins, J. (1995). The influence of gender on physician practice style. *Med Care*, 33(4):407–416.

Berthold, H., Gouni-Berthold, I., Bestehorn, K., Böhm, M., and Krone, W. (2008). Physician gender is associated with the quality of type 2 diabetes care. *Journal of Internal Medicine*, 264(4):340–350.

Berwick, D. and Hackbarth, A. (2012). Eliminating waste in us health care. *JAMA*, 307(14):1513–1516.

Boksem, M. and Tops, M. (2008). Mental fatigue: Costs and benefits. *Brain Research Reviews*, 59:125–139.

Boudreau, J., Hopp, W., McClain, J., and Thomas, L. (2003). On the interface between operations and human resources management. *Manufacturing Service Operations Management*, 5(3):179–202.

Bugter Maessen, A., Winkens, R., Grol, R., Knottnerus, J., Kester, A., Beusmans, G., and Pop, P. (1996). Factors predicting differences among general practitioners in test ordering behaviour and in the response to feedback on test requests. *Family Practice*, 13(3):254–258.

Cameron, A. and Miller, D. (2015). A practitioner's guide to cluster- robust inference. *Journal of Human Resources*, 50(2):317–372.

Cameron, A. and Trivedi, P. (2009). *Microeconometrics Using Stata*. A Stata Press Publication, StataCorp LP, College Station, Texas.

Capilheira, M. and Santos, I. (2006). Population-based study of the epidemiology of diagnostic test ordering. *Journal of Internal Medicine*, 40(2):289–297.

Chan, D. (2018). The efficiency of slacking off: Evidence from the emergency department. *Econometrica*, 86(3):997–1030.

Charness, G. and Gneezy, U. (2012). Strong evidence for gender differences in risk taking. *Journal of Economic Behavior and Organization*, 83(1):50–58.

Choudhry, N., Fletcher, R., and Soumerai, S. (2005). Systematic review: the relationship between clinical experience and quality of health care. *Annals of Internal Medicine*, 142:260–273.

Dafny, L. (2005). How do hospitals respond to price changes? *American Economic Review*, 95(5):1525–1547.

Danziger, S., Levav, J., and Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *PNAS*, 108(17):6889–6892.

de Gracia Gomis, M., Pérez Royo, A., Hernández Aguado, I., Berbegal, J., and Arrese, R. (1999). An analysis of the demand for laboratory tests from primary care in a health area. *Aten Primaria*, 23(1):26–31.

Delasay, M., Ingolfsson, A., Kolfal, B., and Schultz, K. (2016). The influence of load on service times. *Working paper, Carnegie Mellon University, Pittsburgh.*

Deo, S. and Jain, A. (2018). Slow first, fast later: Temporal speed-up in service episodes of finite duration. *Production and Operations Management*, 0(0):1–21.

Dugdale, D., Epstein, R., and Pantilat, S. (1999). Time and the patient–physician relationship. *Journal of General Internal Medicine*, 14(Suppl 1):S34–S40.

Epstein, A., Begg, C., and Mcneil, B. (1984). The effects of physicians' training and personality on test ordering for ambulatory patients. *American Journal of Public Health*, 74(11):1271–1273.

Ferguson, E., James, D., and Madeley, L. (2002). Factors associated with success in medical school: systematic review of the literature. *BMJ*, 324(7343):952–957.

Franks, P., Williams, G., Zwanziger, J., Mooney, C., and Sorbero, M. (2000). Why do physicians vary so widely in their referral rates? *Journal of General Internal Medicine*, 15:163–168.

Freeman, M., Savva, N., and Scholtes, S. (2014). Decomposing the effect of workload on patient outcomes: An empirical analysis of a maternity unit. *Working paper, Cambridge Judge Business School, Cambridge, UK*.

Freeman, M., Savva, N., and Scholtes, S. (2017). Gatekeepers at work: An empirical analysis of a maternity unit. *Management Science*, 63(10):3147–3167.

Fryer, A. and Smellie, W. (2013). Managing demand for laboratory tests: a laboratory toolkit. *Journal of Clinical Pathology*, 66:62–72.

Gaba, D. M. and Howard, S. K. (2002). Fatigue among clinicians and the safety of patients. *New England Journal of Medicine*, 347(16):1249–1255.

Gawande, A. (2009). The cost conundrum: What a texas town can teach us about health care. *The New Yorker*, pages 36–44.

Gong, Q. (2017). Physician learning and treatment choices: Evidence from brain aneurysms. *Working Paper*.

Hartley, R., Charlton, J., Harris, C., and Jarman, B. (1987). Pattern of physicians' use of medical resources in ambulatory settings. *American Journal of Public Health*, 77(5):565–567.

Hockey, G. and Earle, F. (2006). Control over the scheduling of simulated office work reduces the impact of workload on mental fatigue and task performance. *Journal of Experimental Psychology: Applied*, 12(1):50–65.

Holding, D. (1983). *Fatigue*. New York: Wiley.

Hopp, W., Iravani, S., and G.Y., Y. (2007). Operations systems with discretionary task completion. *Management Science*, 53(1):61–77.

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

Huesch, M. (2009). Learning by doing, scale effects, or neither? cardiac surgeons after residency. *Health Services Research*, 44(6):1960–1982.

Hussey, P., Eibner, C., Ridgely, M., and McGlynn, E. (2009). Controlling u.s. health care spending — separating promising from unpromising approaches. *New England Journal of Medicine*, 361(22):2109–2111.

Jerant A, Bertakis KD, F. J. F. P. (2013). Gender of physician as the usual source of care and patient health care utilization and mortality. *Journal of American Board of Family Medicine*, 26(2):138–148.

Kc, D. and Terwiesch, C. (2009). Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science*, 55(9):1486–1498.

Kc, D. and Terwiesch, C. (2012). An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing and Service Operations Management*, 14(1):50–65.

Khalifa, M. and Khalid, P. (2014). Reducing unnecessary laboratory testing using health informatics applications: A case study on a tertiary care hospital. *Procedia Computer Science*, 37:253–260.

Kim, C., McEwen, L., Gerzoff, R., Marrero, D., Mangione, C., Selby, J., and Herman, W. (2005). Is physician gender associated with the quality of diabetes care? *Diabetes Care*, 28(7):1594–1598.

Kuntz, L., Mennicken, R., and Scholtes, S. (2015). Stress on the ward: Evidence of safety tipping points in hospitals. *Management Science*, 61(4):754–771.

Laker, L., Froehle, C., Windeler, J., and Lindsell, C. (2017). Quality and efficiency of the clinical decision-making process: Information overload

and emphasis framing. *Production and Operations Management,forthcoming*, https://doi.org/10.1111/poms.12777.

Leurquin, P., Van Casteren, V., and De Maeseneer, J. (1995). Use of blood tests in general practice: a collaborative study in eight european countries. *British Journal of General Practice*, 45(390):21–25.

Liu, J., Ma, J., Wang, J., Zeng, D., Song, H., and Wang, L.and Cao, Z. (2016). Comorbidity analysis according to sex and age in hypertension patients in china. *International Journal of Medical Sciences*, 13(2):99–107.

Martins, C., Godycki-Cwirko, M., Heleno, B., and BrodersenChan, J. (2018). Quaternary prevention: reviewing the concept. *European Journal of General Practice*, 24(1):106–111.

Mazur, L., Mosaly, P., Moore, C., Comitz, E., Yu, F., Falchook, A., Eblan, M., Hoyle, L., Tracton, G., Chera, B., and Marks, L. (2016). Toward a better understanding of task demands, workload, and performance during physician-computer interactions. *Journal of American Medical Informatics Association*, 23:1113–1120.

Miyakis, S., Karamanof, G., and Liontos, M.and Mountokalakis, T. (2006). Factors contributing to inappropriate ordering of tests in an academic medical department and the effect of an educational feedback strategy. *Postgrad Med. J.*, 82:823–829.

Montgomery, D., Peck, E., and Vining, G. (2001). *Introduction to Linear Regression Analysis.* Wiley.

Oliva, R. and Sterman, J. (2001). Cutting corners and working overtime: Quality erosion in the service industry. *Management Science*, 47(7):894–914.

Piccirillo, J., Vlahiotis, A., Barrett, L., Flood, K., Spitznagel, E., and Steyerberg, E. (2008). The changing prevalence of comorbidity across the age spectrum. *Critical Reviews in Oncology/Hematology*, 67(2):124–132.

Powell, A., Savin, S., and Savva, N. (2012). Physician workload and hospital reimbursement: Overworked physicians generate less revenue per patient. *Manufacturing and Service Operations Management*, 14(4):512–528.

Powell, M. and Ansic, D. (1997). Gender differences in risk behaviour in financial decision-making: an experimental analysis. *Journal of Economic Psychology*, 18:605–628.

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

Rosenbaum, L. (2017). The less-is-more crusade — are we overmedicalizing or oversimplifying? *The New England Journal of Medicine*, 377(24):1042–1062.

Roter, D. and Hall, J. (2004). Physician gender and patient centered communication: a critical review of empirical research. *Annual Review of Public Health*, 25:497–519.

Roter, D., Hall, J., and Aoki, Y. (2002). Physician gender effects in medical communication: a meta-analytic review. *JAMA*, 288:756–764.

Salloum, S. and Franssen, E. (1993). Laboratory investigations in general practice. *Canadian Family Physician*, 39:1055–1061.

Schultz, K., Juran, D., and Boudreau, J. (1999). The effects of low inventory on the development of productivity norms. *Management Science*, 45(12):1664–1678.

Schultz, K., Juran, D., Boudreau, J., J.O., M., and Thomas, L. (1998). Modeling and worker motivation in jit production systems. *Management Science*, 44(12):1595–1607.

Shye, D., Freeborn, D., Romeo, J., and Eraker, S. (1998). Understanding physicians' imaging test use in low back pain care: the role of focus groups. *International Journal for Quality in Health Care*, 10(2):83–91.

Silverman, E. and Skinner, J. (2004). Medicare upcoding and hospital ownership. *Journal of Health Economics*, 23(2):369–389.

Streja, D. and Rabkin, S. (1999). Factors associated with implementation of preventive care measures in patients with diabetes mellitus. *Archives of Internal Medicine*, 159:294–302.

Sun, B., Adams, J., Orav, E., Rucker, D., Brennan, T., and Burstin, H. (2000). Determinants of patient satisfaction and willingness to return with emergency care. *Annals of Emergency Medicine*, 35(5):3426–434.

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

Tanabe, S. and Nishihara, S. (2004). Productivity and fatigue. *Indoor Air*, 14(Suppl. 7):126–133.

Tsugawa, Y., Jena, A., Figueroa, J., Orav, E., Blumenthal, D., and Jha, A. (2017a). Comparison of hospital mortality and readmission rates for medicare patients treated by male vs female physicians. *JAMA Internal Medicine*, 177(2):206–213.

Tsugawa, Y., Jena, A., Orav, E., Blumenthal, D., Tsai, T., Mehtsun, W., and Jha, A. (2018). Age and sex of surgeons and mortality of older surgical patients: observational study. *BMJ*, 361:k1343.

Tsugawa, Y., Newhouse, J., Zaslavsky, A., Blumenthal, D., and Jena, A. (2017b). Physician age and outcomes in elderly patients in hospital in the us: observational study. *BMJ*, 357:j1797.

Van der Linden, D., Frese, M., and T.F., M. (2003). Mental fatigue and the control of cognitive processes: Effects on perseveration and planning. *Acta Psychologica*, 113:981–989.

Van der Weijden, T., van Bokhoven, M., Dinant, G., van Hasselt, C., and Grol, R. (2002). Understanding laboratory testing in diagnostic uncertainty: a qualitative study in general practice. *British Journal of General Practice*, 52(485):974–980.

Verstappen, W., ter Riet, G., Dubois, W., Winkens, R., Grol, R., and van der Weijden, T. (2004). Variation in test ordering behaviour of gps: professional or context-related factors? *Family Practice*, 21(4):387–395.

Vinker, S., Kvint, I., Erez, R., Elhayany, A., and Kahan, E. (2007). Effect of the characteristics of family physicians on their utilisation of laboratory tests. *British Journal of General Practice*, 57:377–382.

Wallis, C., Ravi, B., Coburn, N., Nam, R., Detsky, A., and Satkunasivam, R. (2017). Comparison of postoperative outcomes among patients treated by male and female surgeons: a population based matched cohort study. *BMJ*, 359:j4366.

Whiting, P., Toerien, M., de Salis, I., Sterne, J., Dieppe, P., Egger, M., and Fahey, T. (2007). A review identifies and classifies reasons for ordering diagnostic tests. *Journal of Clinical Epidemiology*, 60:981–989.

Wooldridge, J. (2009). *Introductory Econometrics: A Modern Approach.* South-Western Cengage Learning.

Zhi, M., Ding, E., Theisen-Toupal, J., Whelan, J., and Arnaout, R. (2013). The landscape of inappropriate laboratory testing: A 15-year meta-analysis. *PLoS*, ONE8(11):e78962.

Appendix A

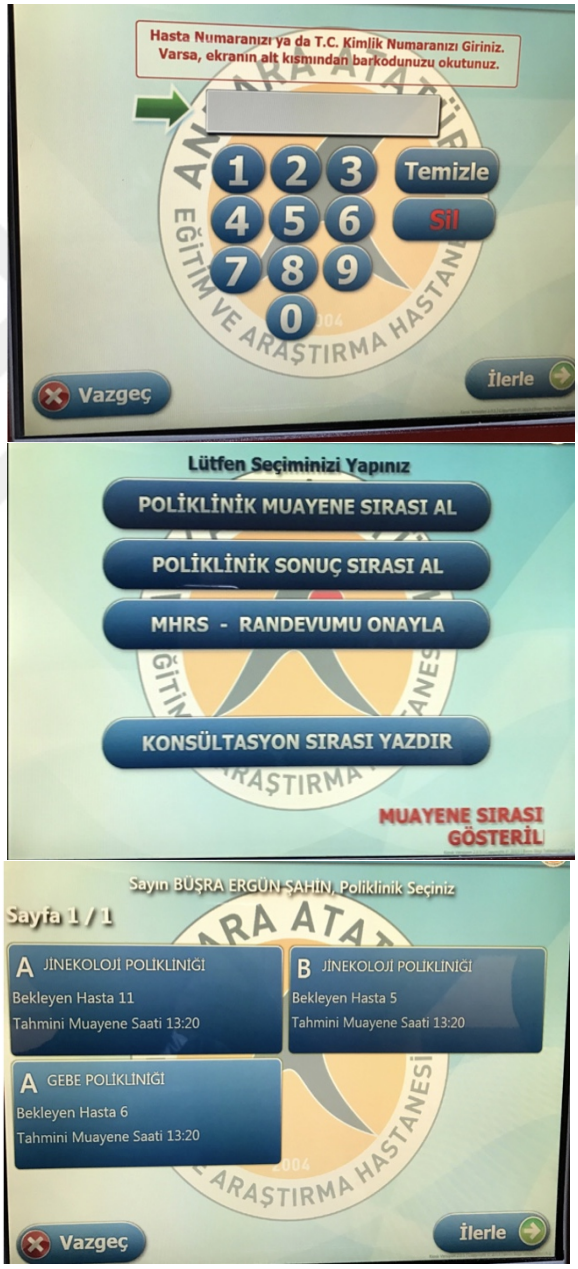


Figure A.1: Patient registration booth

Figure A.2: Waiting area counters to follow examination sequence number

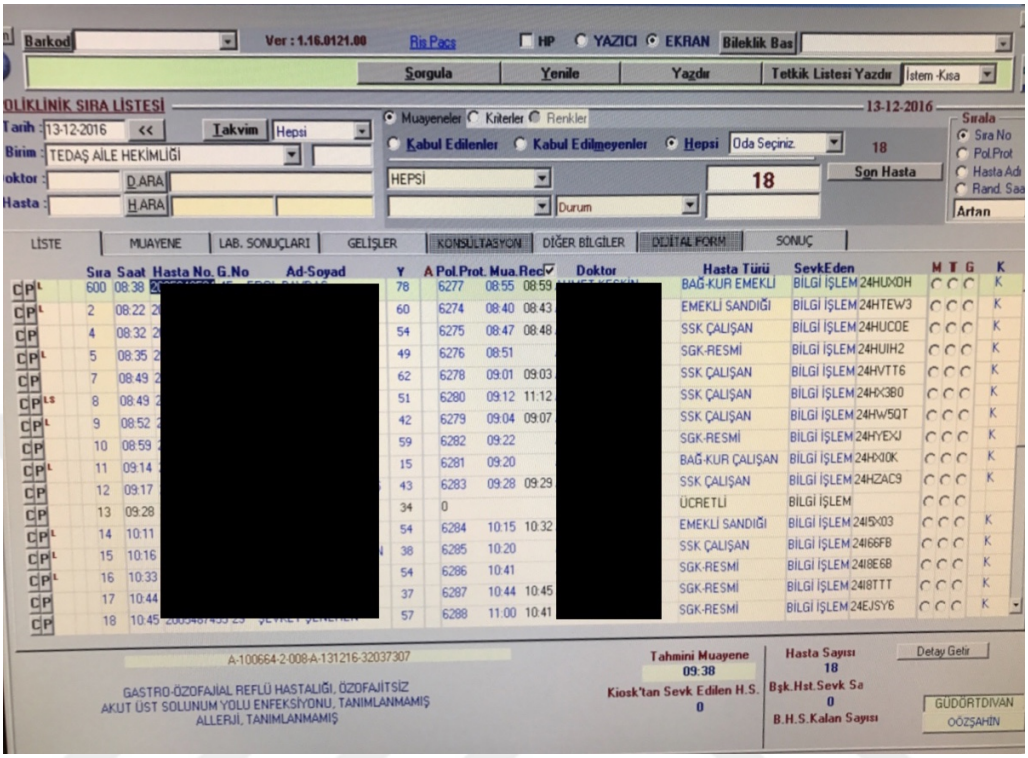


Figure A.3: Physician's main page on their computer with the list of registered patients

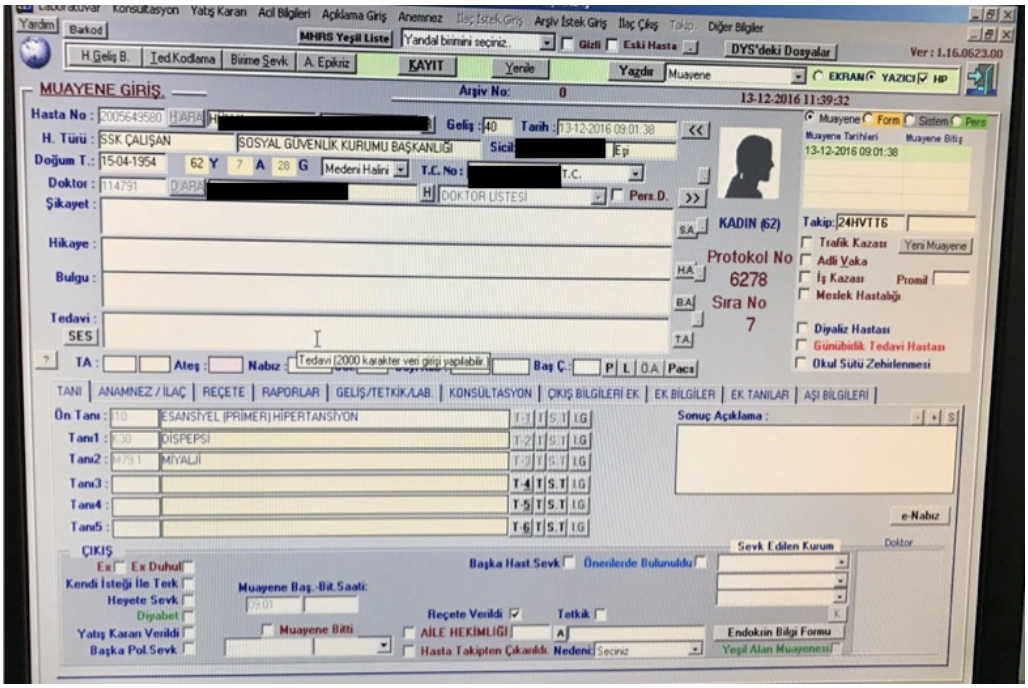


Figure A.4: Patient's page for electronic medical records

Appendix B

Table B.1: 20 most frequently requested diagnostic tests

Name of Diagnostic Test	**TOTAL**	**%TOTAL**
CHOLESTEROL	5697	4.07%
TRIGLYCERIDE	5665	4.05%
HDL CHOLESTEROL	5508	3.94%
GLUCOSE	5143	3.68%
CREATININ	5034	3.60%
ALANINE AMINOTRANSFERASE (ALT)	4994	3.57%
IRON (SERUM)	4865	3.48%
HEMOGRAM	4799	3.43%
UREA	4794	3.43%
ASPARTAT TRANSAMINASE (AST)	4581	3.27%
TSH	4255	73.04%
SODIUM (NA)	4182	2.99%
POTASIUM	4179	2.99%
VITAMIN B12	3795	2.71%
LDL CHOLESTEROL	3793	2.71%
FERRITIN	3705	2.65%
EGFR	3522	2.52%
GAMMA GLUTAMIL TRANSFERASE (GGT)	3488	2.49%
FOLATE	3345	2.39%
CALSIUM(CA)	3206	2.29%

Table B.2: 10 most frequent ICD codes

ICD Code	**Description**	**Quantity**	**Percentage**
Z04.8	Encounter for examination and observation for other specified reasons	32,541	74%
E13.8	Other specified diabetes mellitus with unspecified complications	2,390	5.44%
I10	Essential (primary) hypertension	1,116	2.54%
E03.9	Hypothyroidism, unspecified	867	1.97%
Z00.8	Encounter for other general examination	840	1.91%
K30	Functional dyspepsia	573	1.30%
K21.9	Gastro-esophageal reflux disease without esophagitis	377	0.86%
M25.5	Pain in joint	366	0.83%
D64.9	Malaise and fatigue	366	0.83%
R10.4	Other and unspecified abdominal pain	349	0.79%

Appendix C

Table C.1: Correlation table

n=11,271	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
ntests(1)	1.000										
waiting(2)	0.1557	1.000									
finishedload(3)	-0.1973	-0.4893	1.000								
time(4)	-0.2139	-0.6132	0.7781	1.000							
status(5)	0.2665	0.0707	-0.0843	-0.0903	1.000						
age(6)	0.0788	0.1343	-0.1483	-0.1320	0.0109	1.000					
gender(7)	0.0057	0.0041	0.0239	0.0274	-0.0265	-0.0004	1.000				
year(8)	0.0004	-0.0742	-0.0504	0.0425	-0.0420	-0.0483	0.0267	1.000			
month(9)	0.0541	-0.0705	0.0031	0.0514	-0.0292	-0.0325	0.0229	-0.0034	1.000		
day(10)	0.0489	0.0583	0.0067	0.0011	0.0255	0.0075	-0.0030	0.0164	0.0423	1.000	
nvisits(11)	-0.0268	0.0314	-0.0654	-0.0591	0.0095	0.2316	0.0734	0.0001	0.0194	-0.0076	1.000

Table C.2: Results from the ZINB model with control variables

Variables	Negative Binomial	Inflate (Logit)	Marginal Effects
waiting	−0.002*	−0.029***	0.082***
	(0.001)	(0.003)	
finishedload	−0.004***	0.002	−0.050***
	(0.001)	(0.002)	
time	−0.055***	0.234***	−1.512***
	(0.017)	(0.037)	
year	−0.039	−0.585***	1.737***
	(0.024)	(0.052)	
month	0.003	−0.027***	0.139***
	(0.003)	(0.006)	
day	0.010	−0.042***	0.269***
	(0.007)	(0.016)	
age	0.004***	0.004***	0.026***
	(0.001)	(0.002)	
gender	0.039*	−0.133***	0.946***
	(0.020)	(0.044)	
nvisits	−0.001***	0.004***	−0.024***
	(0.000)	(0.001)	
status(*new*)	0.818***	−2.055***	10.893***
	(0.052)	(0.077)	
2.*physician*	0.241***	0.106	1.941**
	(0.059)	(0.137)	
3.*physician*	0.224***	−0.021	2.231***
	(0.042)	(0.097)	
4.*physician*	0.200***	−0.005	1.911***
	(0.049)	(0.101)	
5.*physician*	0.363***	−0.148	4.352***
	(0.052)	(0.125)	
6.*physician*	0.366***	−0.148**	4.395***
	(0.029)	(0.062)	
ICD codes are not shown due to space limitations			
Observations	11,271		
Standard errors in parentheses			
*** p<0.01, ** p<0.05, * p<0.1			

Appendix D

Table D.1: Results from the random effects logit model

Variables	Random Effects Logit Model
waiting	0.024***
	(0.004)
finishedload	−0.005*
	(0.003)
rho	0.161
	(0.035)
Observations	11,271
Number of patients	9,321
Likelihood-ratio test of rho=0: chibar2(01)=28.58	
Prob≥chibar2 = 0.000	
Standard errors in parentheses	
*** p<0.01, ** p<0.05, * p<0.1	

Table D.2: Comparison of results from the random effects logit model and logit model

Variable	Random Effects Logit	Logit
waiting	0.024^{***}	0.023^{***}
	(0.004)	(0.003)
finishedload	-0.005^{*}	-0.004
	(0.003)	(0.002)
Observations	11,271	11,210
Number of patients	9,321	
Standard errors in parentheses		
*** p<0.01, ** p<0.05, * p<0.1		

Table D.3: Results of the ZINB model with robust and clustered robust standard errors

Variables	Main	Robust	Clustered Robust
Negative Binomial			
waiting	-0.002^{*}	-0.002^{**}	-0.002^{**}
	(0.001)	(0.001)	(0.001)
finishedload	-0.004^{***}	-0.004^{***}	-0.004^{***}
	(0.001)	(0.001)	(0.001)
Logit(Inflate)			
waiting	-0.029^{***}	-0.029^{***}	-0.029^{***}
	(0.003)	(0.003)	(0.003)
finishedload	0.002	0.002	0.002
	(0.002)	(0.002)	(0.002)
Marginal Effects			
waiting	0.082^{***}	0.082^{***}	0.082^{***}
finishedload	-0.050^{***}	-0.050^{***}	-0.050^{***}
Observations		11,271	
Standard errors in parentheses			
*** p<0.01, ** p<0.05, * p<0.1			

Table D.4: Correlation table for instrumental variables

n=10,259	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
avwaiting(12)	0.2087	0.6700	-0.7397	-0.9202	0.1016	0.1324	-0.0191	-0.0178	-0.0442	0.0736	0.0517
reg_interval(13)	-0.1585	-0.5734	0.5323	0.6450	-0.0747	-0.0511	0.0063	0.1927	0.0973	-0.0167	-0.0374
	(12)	(13)									
avwaiting(12)	1.000										
reg_interval(13)	-0.6333	1.000									

ntests(1), *waiting*(2), *finishedload*(3), *time*(4), *status*(5), *age*(6), *gender*(7), *year*(8), *month*(9), *day*(10)

Table D.5: Results of endogeneity test for waiting

Number of obs = 10,259

$F(138, 10120) = 19.44$

$Prob > F = 0.0000$

Total (centered) SS = 1906885.693 — Centered R2 = 0.2095

Total (uncentered) SS = 3391607 — Uncentered R2 = 0.5556

Residual SS = 1507354.87 — Root MSE = 12.12

ntests	Coef.	Std. Err.	z	P> $\|z\|$	[95% Conf. Interval]	
waiting	0.062	0.043	1.42	0.157	-0.024	0.147
finishedload	-0.037	0.013	-2.79	0.005	-0.063	-0.011

Underidentification test (Anderson canon. corr. LM statistic):	1696.049
Chi-sq(2) P-val =	0.0000
Weak identification test (Cragg-Donald Wald F statistic):	1002.126
Stock-Yogo weak ID test critical values: 10% maximal IV size	19.93
15% maximal IV size	11.59
20% maximal IV size	8.75
25% maximal IV size	7.25
Sargan statistic (overidentification test of all instruments):	0.626
Sargan statistic ($Chi-sq(1) P-val =$	0.4289

Table D.6: Results from structural model approach for endogeneity

Variables	IV MODEL	NB MODEL
waiting	0.010*	0.007***
	(0.006)	(0.002)
finishedload	−0.005**	−0.006***
	(0.002)	(0.002)
$\hat{v}_i$	−0.004	
	(0.006)	
Observations	10,259	11,271
Standard errors in parentheses		
*** p<0.01, ** p<0.05, * p<0.1		

Table D.7: Results from the ZINB model with continuous time variable

Variables	Negative Binomial	Inflate (Logit)	Marginal Effects
waiting	−0.002	−0.023***	0.068***
	(0.001)	(0.003)	
timecontinuous	−0.001***	0.005***	−0.029***
	(0.000)	(0.001)	
finishedload	−0.003**	0.000	−0.037**
	(0.001)	(0.003)	
Observations		11,271	
Standard errors in parentheses			
*** p<0.01, ** p<0.05, * p<0.1			

Table D.8: Results from the ZINB model with alternative finishedload variable

Variables	Negative Binomial	Inflate (Logit)	Marginal Effects
waiting	−0.002	−0.029***	0.092***
	(0.001)	(0.003)	
finload1hour	−0.005***	−0.004	−0.039*
	(0.001)	(0.004)	
Observations		11,271	
Standard errors in parentheses			
*** p<0.01, ** p<0.05, * p<0.1			

Table D.9: Results from logit models for each time interval

Variables	time0	time1	time2	time3	time4
waiting	0.001	0.013**	0.031***	0.038***	0.056*
	(0.016)	(0.005)	(0.006)	(0.010)	(0.033)
finishedload	0.067	−0.005	−0.010**	−0.005	0.008
	(0.092)	(0.007)	(0.004)	(0.004)	(0.007)
Observations	416	3,591	3,537	2,591	838
Standard errors in parentheses					
*** p<0.01, ** p<0.05, * p<0.1					

Table D.10: Results from the ZINB model with alternative load variable

Variables	Negative Binomial	Inflate (Logit)	Marginal Effects
remainingload	−0.322	−4.160***	11.923***
	(0.212)	(0.493)	
finishedload	−0.004***	0.004*	−0.056***
	(0.001)	(0.002)	
Observations		11,271	
Standard errors in parentheses			
*** p<0.01, ** p<0.05, * p<0.1			

Table D.11: Results from the ZINB model with interaction term

Variables	Negative Binomial	Inflate (Logit)	Marginal Effects
waiting	−0.001	0.008	0.096***
	(0.002)	(0.006)	
time	−0.050**	0.316***	−0.907***
	(0.020)	(0.041)	
c.waiting#c.time	−0.001	−0.021***	
	(0.001)	(0.002)	
finishedload	−0.004***	0.007***	−0.070***
	(0.001)	(0.002)	
Observations		11,271	
Standard errors in parentheses			
*** p<0.01, ** p<0.05, * p<0.1			

Table D.12: Results from the ZINB model with normalized waiting variable

Variables	Negative Binomial	Inflate (Logit)	Marginal Effects
normwaiting	-0.019^{*}	-0.249^{***}	0.708^{***}
	(0.011)	(0.028)	
finishedload	-0.004^{***}	0.002	-0.050^{***}
	(0.001)	(0.002)	
Observations		11,271	
Standard errors in parentheses			
*** p<0.01, ** p<0.05, * p<0.1			

Table D.13: Results from the ZINB model with alternative samples

Variables	Hypertension	Diabetes	Z04.8
Negative Binomial			
waiting	−0.0026	−0.0026	0.0005
	(0.0030)	(0.0020)	(0.0007)
finishedload	−0.0003	-0.0038^{**}	−0.0008
	(0.0027)	(0.0018)	(0.0005)
Logit(Inflate)			
waiting	-0.0394^{***}	-0.0362^{***}	-0.0164^{***}
	(0.0091)	(0.0074)	(0.0023)
finishedload	0.0079	0.0041	0.0055^{***}
	(0.0067)	(0.0054)	(0.0014)
Marginal Effects			
waiting	0.1620^{***}	0.1060^{**}	0.0574^{***}
finishedload	−0.0407	-0.0739^{**}	-0.0287^{***}
Observations	1,116	2,390	32,541
Standard errors in parentheses			
*** p<0.01, ** p<0.05, * p<0.1			

Table D.14: Results from the ZINB model with all the ICD Codes

Variables	200 Iterations	300 Iterations	500 Iterations	1000 Iterations
Negative Binomial				
waiting	−0.0014**	−0.0012**	−0.0007	−0.0001
	(0.0006)	(0.0006)	(0.0006)	(0.0006)
finishedload	−0.0013***	−0.0010**	−0.0009*	−0.0009*
	(0.0005)	(0.0005)	(0.0005)	(0.0005)
Logit(Inflate)				
waiting	−0.0184***	−0.0185***	−0.0185***	−0.0184***
	(0.0018)	(0.0018)	(0.0018)	(0.0018)
finishedload	0.0034***	0.0035***	0.0035***	0.0035***
	(0.0012)	(0.0012)	(0.0012)	(0.0012)
Marginal Effects				
waiting	0.0412***	0.0440***	0.0508***	0.0585***
finishedload	−0.0287***	−0.0252***	−0.0234***	−0.0237***
Observations	43,966			
Standard errors in parentheses				
*** p<0.01, ** p<0.05, * p<0.1				

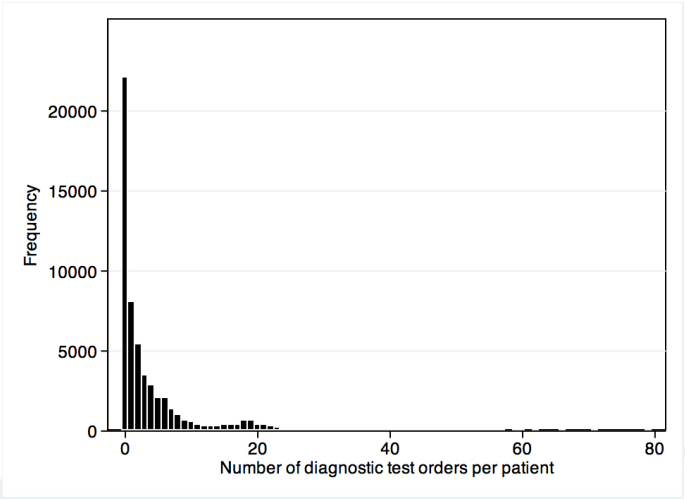


Figure D.1: Histogram of the number of diagnostic test orders per patient in the gynecology polyclinic

Table D.15: Results from the ZINB model with gynecology polyclinic data

Variables	Negative Binomial	Inflate (Logit)	Marginal Effects
waiting	-0.003^{***}	-0.107^{***}	0.010^{**}
	(0.001)	(0.016)	
finishedload	-0.003^{***}	0.009^{***}	-0.012^{***}
	(0.001)	(0.003)	
Observations		55,424	
	Standard errors in parentheses		
	*** p<0.01, ** p<0.05, * p<0.1		

Appendix E

Table E.1: Number of patients seen by each physician at each location by specialty

PhysicianID	Main	District 1	District 2	District 3	District 4	Total
	Internal Medicine					
113727		7,835				7,835
114985			7,412			7,412
115028	6,447					6,447
113481	6,064	29				6,093
113323		184	5,498			5,682
100874				4,243		4,243
115310	3,823					3,823
115328		3,335				3,335
113854					3,305	3,305
115103	2,522					2,522
115266	2,170					2,170
114952	618					618
Total						53,485
	Gynecology					
100127	6,471	79		92		6,642
100313	3,073	381		283		3,737
100141	2,497	617		422		3,536
111903	3,419	2		9		3,430
113939	292	18		33		343
114431	3,022	459		214		3,695
112149	2,876	260		219		3,355
112282	2,902	287		163		3,352
115099	2,727	340		107		3,174
113407	1,983	321		149		2,453
113271				2,001		2,001
114433	1,672	239		76		1,987
115311					1,265	1,265
114625	975	107		39		1,121
Total						40,091
	Pediatrics					
115692	695	27		34		756
100525	351	12		44		407
115655	148	37		63		248
101343	1,842	37		261		2,140
101344	1,832	62		210		2,104
101219	1,451	102		317		1,870
114911	1,364	103		316		1,783
100459	1,220	104		334		1,658
115053	1,412	41		153		1,606
113622	882	78		150		1,110
Total						13,682
	Endocrinology					
100161	1,405					1,405
101033	593					593
Total						1,998
	Hematology					
112370	2,819					2,819
112760	3,432					3,432
113747	3,237					3,237
115441	510					510
Total						9,998
Total	76,746	15,096	12,910	9,932	4,570	119,254

Table E.2: 10 most frequent ICD codes

ICD Code	**Description**	**Percentage**
Z00.8	Encounter for other general examination	15.20%
Z04.8	Encounter for examination and observation for other specified reasons	13.20%
Z33	Pregnant state, incidental	10.40%
N92.6	Irregular menstruation, unspecified	7.20%
R10.2	Pelvic and perineal pain	6.90%
D64.9	Anemia, unspecified	3.50%
R53	Malaise and fatigue	2.80%
N76.0	Acute vaginitis	2.60%
I10	Essential (primary) hypertension	1.90%
D50.9	Iron deficiency anemia, unspecified	1.80%

Table E.3: Results of negative binomial model with random effects

Variables	XTNBREG
physicianexperience	1.005***
	(0.001)
physiciangender : *female*	1.104***
	(0.014)
patientgender : *female*	1.040***
	(0.009)
waiting	1.005***
	(0.001)
finishedload	0.997***
	(0.000)
time	0.839***
	(0.005)
noofvisits	0.999***
	(0.000)
status : *new*	3.495***
	(0.043)
dayofweek	0.983***
	(0.002)
month	1.008***
	(0.001)
patientage	1.000
	(0.000)
mainhospital	0.955**
	(0.017)
*district*2	0.766***
	(0.017)
*district*3	1.176***
	(0.026)
*district*4	0.719***
	(0.022)
speciality : *hematology*	0.876**
	(0.056)
speciality : *internal*	0.942
	(0.057)
speciality : *gynecology*	0.787***
	(0.059)
speciality : *pediatrics*	0.632***
	(0.040)
ICD codes are not shown due to space limitations	
Observations	119,254
# physicians	42
AIC	679,053.2
Log likelihood	−339,269.59
Standard errors in parentheses	
*** p<0.01, ** p<0.05, * p<0.1	

Table E.4: Practice variations of physicians at different work locations

Variables	Pediatrics	Pediatrics and Gynecology
physicianexperience	1.014***	1.007***
	(0.002)	(0.001)
physiciangender : *female*		1.084***
		(0.023)
patientgender : *female*	1.005	1.029
	(0.022)	(0.023)
mainhospital	0.963	0.994
	(0.058)	(0.025)
*district*3	0.977	0.993
	(0.063)	(0.030)
speciality : *pediatrics*		1.125*
		(0.075)
Observations	13,682	50,507
# physicians	10	22
Standard errors in parentheses		
*** p<0.01, ** p<0.05, * p<0.1		

Table E.5: Alternative model with experience of physician in last 6 month

Variables	Alt. Experience Model
*experience*6*month*	1.000***
	(0.000)
physiciangender : *female*	1.226**
	(0.111)
Observations	35,636
# physicians	23
Standard errors in parentheses	
*** p<0.01, ** p<0.05, * p<0.1	

Table E.6: Marginal effects of skewness model for physicians' gender and experience

Variables	Marginal Effects
physicianexperience	
0 years	0.362
5 years	0.348
10 years	0.335
15 years	0.321
20 years	0.308
25 years	0.294
30 years	0.281
35 years	0.267
physiciangender	
$female$	0.295
male	0.357

Table E.7: Results of skewness model for distribution of daily load more than 10, 15 and 20

Variables	Load>10	Load>15	Load>20
physiciangender : $female$	-0.063	-0.07^{**}	-0.044
	(0.050)	(0.034)	(0.047)
physicianexperience	-0.003	-0.003^{*}	-0.002
	(0.002)	(0.002)	(0.002)
initialload	0.019^{***}	0.015^{***}	0.011^{***}
	(0.002)	(0.002)	(0.002)
dailyload	-0.011^{***}	-0.009^{***}	-0.007^{***}
	(0.001)	(0.001)	(0.001)
Observations	3,832	3,183	2,535
R-squared	0.092	0.100	0.108
# physicians	42	41	39

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table E.8: Results of chi-square model for distribution of daily load more than 10, 15 and 20

Variables	Load>10	Load>15	Load>20
$physiciangender : female$	−0.558	−0.680*	−0.590
	(0.402)	(0.412)	(0.564)
physicianexperience	−0.070***	−0.082***	−0.086***
	(0.017)	(0.018)	(0.027)
initialload	0.229***	0.219***	0.206***
	(0.022)	(0.024)	(0.027)
dailyload	−0.031***	−0.040***	−0.028*
	(0.012)	(0.013)	(0.016)
Observations	3,832	3,183	2,535
R-squared	0.178	0.184	0.194
# physicians	42	41	39
Standard errors in parentheses			
*** p<0.01, ** p<0.05, * p<0.1			